Verifying Forecasts Spatially

Eric Gilleland
Research Applications Laboratory, the National Center for Atmospheric Research, Boulder, Colorado

David A. Ahijevych
Mesoscale and Microscale Meteorology, the National Center for Atmospheric Research, Boulder, Colorado

Barbara G. Brown
Research Applications Laboratory, the National Center for Atmospheric Research, Boulder, Colorado

Elizabeth E. Ebert
Centre for Australian Weather and Climate Research,
Melbourne, Victoria, Australia

Corresponding author address: Eric Gilleland, Research Applications Laboratory, the National Center for Atmospheric Research, 3450 Mitchell Lane,
Boulder, Colorado 80301.
E-mail: EricG@ucar.edu

Abstract

Numerous new methods have been proposed for using spatial information to better quantify and diagnose forecast performance when forecasts and observations are both available on the same grid. Gilleland *et al.* (2009a) classify the majority of the new spatial verification methods into four broad categories (neighborhood, scale separation, features-based, and field deformation), which themselves can be further generalized into two categories of filter and displacement. Because the methods make use of spatial information in widely different ways, users may be uncertain about what types of information each provides, and which methods may be most beneficial for particular applications. An international project, the Spatial Forecast Verification Methods Inter-Comparison Project (ICP, http://www.ral.ucar.edu/projects/icp), was formed to address these questions. This project was coordinated by NCAR and facilitated by the WMO/WWRP Joint Working Group on Forecast Verification Research. An overview of the methods involved in the project is provided here with some initial guidelines about when each of the verification approaches may be most appropriate.

Capsule

New spatial forecast methods are put to the test in an international effort to provide information about how each method handles different types of forecast errors and what information about forecast performance is provided.

1. Introduction

Verification of a forecast field presents many challenges, especially at higher resolutions. When assessing forecast performance at a single point, straightforward summary statistics (e.g., root mean square error, RMSE) are meaningful because they give an intuitive notion of how well the forecasts matched the observations at that point. It is also straightforward to identify hits, misses, false alarms, and correct negatives; all of which give rise to numerous useful summary statistics (e.g., probability of detection, POD; Gilbert Skill Score, GSS) and diagnostics (e.g., relative operating characteristic) of forecast performance (see, e.g., Jolliffe and Stephenson, 2003; Wilks, 1995, for more on traditional verification). However, when interest is in spatially-coherent structures, these notions are not as simple to determine. Further, several new types of errors become relevant, which bring about a new set of verification questions. Are there spatial displacement errors? Does the scale-dependent variance of the forecast field match the spatial structure that was observed? Did the forecast under- or over-predict the spatial extent of a storm system? Are there orientation errors for specific structures in the field? At what scales do the forecasts have skill?

Numerous methods have been proposed in recent years to address these issues, and some are already in use at meteorological centers around the world. Because the methods are relatively new, however, it is not always clear what approaches are best suited for answering particular questions about forecast performance. Which ones should be used for specific purposes, and which provide analogous information? How can uncertainty information be determined about the results? Do methods have any unintuitive characteristics that a user should know about? Such questions provided the impetus for the Spatial Forecast Verification Methods Inter-Comparison project (ICP, http://www.ral.ucar.edu/projects/icp). Although not every approach is currently under study in the ICP, a reasonably representative subset is included; in most cases by the researchers who originally proposed them.

Several test cases have been provided in order to make head-to-head comparisons of how each method addresses forecast performance. So far, three sets of cases focused on quantitative precipitation forecasts (QPF) are being used, with plans to incorporate more varied cases in the future. The cases under study now include three configurations of the Weather Research Forecasting (WRF) model, known perturbations of one of these cases, and simple geometric cases with prescribed spatial displacement and/or spatial extent errors (Fig. 1 shows some example test cases). The geometric cases are pure examples of common forecast errors and provide useful information about the output of each method. Perturbed "real" cases give a better idea of how well the methods assess forecast performance with known errors under a more realistic setting than the geometric cases. To compare the methods with the real WRF cases, a subjective evaluation was carried out, though even subjective evaluations can be varied and even misleading. More detailed information about the test cases and the subjective evaluations can be found in Ahijevych *et al.* (2009). A more complete literature review and qualitative comparison of the methods can be found in Gilleland *et al.* (2009a).

2. The Methods

Most of the techniques proposed can be classified into one of four categories: (a) scale separation (or decomposition), (b) neighborhood (or fuzzy), (c) features-based (or object-based), and (d) field deformation. The first two could be further generalized as filter methods where the scale separation methods take advantage of band-pass filters to separate forecast performance at different *physical* scales, and the neighborhood methods utilize smoothing filters. An advantage of both approaches is the general ability to describe the "scale" at which the forecast attains a particular level of skill. Similarly, the features-based and field deformation approaches could be grouped together as spatial displacement methods, although both can give more information about forecast skill than just spatial displacement.

Figure 2 shows a schematic of the general categories. Traditional scores or other summary statistics for the filter methods are applied at different scales, where in the case of the neighborhood methods, these statistics are calculated for different neighborhood sizes, and for scale separation, they are calculated for different spectral bands that isolate phenomena of a particular size. To illustrate the scale-separation approach, the red image in the upper left corner of the scale-separation window shows the binary image of a large-scale weather system, and constituent wavelet components in the other subpanels. The displacement methods attempt to displace the forecast field spatially to better match the observations. Information about the amount of displacement necessary can be gleaned along with various other diagnostics and summary statistics. The primary difference between the methods is that the features-based approaches identify individual features (or objects) within a field, and analyze these structures separately, whereas the field deformation methods apply to the entire field as a whole.

Not all spatial verification methods fall nicely into the categories mentioned above, though they may share some characteristics. The composite method of Nachamkin (2004, 2009), the multi-scale spatial statistics proposed in Harris *et al.* (2001), the Gaussian mixture models approach of Lakshmanan and Kain (2009), and the variogram approach of Marzban and Sangathe (2009) might be better classified as distribution methods because they make distributional comparisons between the two fields. The Forecast Quality Index method of Venugopal *et al.* (2005) provides summary information involving a spatial displacement metric applied to the entire field but does not attempt to displace the field in any way. Mesinger (2008) introduces a modified GSS that corrects for frequency bias. As a result, the statistic removes the advantage gained by overforecasting the spatial extent of features that are displaced from the observations (e.g., Baldwin and Kain, 2006), and better reflects the error caused by displacement alone.

The following subsections provide brief descriptions of the methods by type.

a. Neighborhood Approaches

Neighborhood approaches differ from one another primarily by the type of smoothing filter applied. The filter is applied to one or both of the forecast and observed fields, and summary statistics (e.g., traditional verification statistics) are applied to the filtered fields. Some neighborhood methods allow verification against point observations. Further, most filters preserve the peak values, which is important for severe wind reports or maximum hail size. Information about the scales at which a forecast attains a desired level of skill can be obtained by iteratively increasing the neighborhood size to which the filter is applied. In this sense, the term “scale” differs from that used in conjunction with the scale separation methods in that here one scale is not independent from another; as the scale increases, the overall field becomes less sharply defined, usually resulting in better skill. Ebert (2008) provides a thorough review of the neighborhood approaches, and the reader is directed there for more information and references. Mittermaier and Roberts (2009) apply the Fractions Skill Score (FSS) of Roberts and Lean (2008) to the ICP test cases, and find it to be a good measure of spatial accuracy in addition to learning about scales of useful skill.

b. Scale Separation Approaches

Scale separation techniques are not new to forecast verification (e.g., Brooks and Levine, 1997; Tustison *et al.*, 2001). Typically, each field is decomposed using some type of band-pass filter (e.g., Fourier, wavelets, etc.), and the two fields are compared using traditional verification techniques at each spectral scale. Note that the term "scale" used here refers to physical features such as separate large-scale frontal systems or smaller scale convective showers. The techniques attempt to assess the scale-dependent error, determine the scales where a forecast has a desired level of skill, and to investigate a forecast's ability to reproduce the observed variability scale structure of the observed field.

Exceptions to the generalizations outlined above include the intensity scale (IS) method of Casati *et al.* (2004) and Casati (2009), the structure function and moment-scale analysis of Harris *et al.* (2001), and the variogram approach of Marzban and Sangathe (2009). The IS approach is fairly close in spirit to the above description, except that the forecast and observation fields are first transformed into binary fields (taken for several different intensity thresholds), the difference between the binary fields is then taken, and the resulting error fields are decomposed using a two-dimensional Haar wavelet. Finally, a traditional skill score based on the mean squared error of these images (Casati, 2009 additionally calculates the energy) is evaluated for each scale component and intensity threshold. The approaches of Harris *et al.* (2001) and Marzban and Sandgathe (2009) do not perform verification at separate scales, and instead investigate the two fields individually so that information about their marginal distributions is given instead of their joint distribution.

Lack *et al.* (2009) utilize Fourier decomposition before applying a version of the features-based technique introduced in Micheas *et al.* (2007). To the best of our knowledge, it is the first method introduced that directly diagnoses the spatial displacement errors (and other errors) for isolated or specific physical scales. Indeed, the method is a features-based method that uses a Fourier decomposition to identify objects. Such a combination across types of methods is very natural, and it is likely that more crossovers will be proposed.

c. Features-based Approaches

These methods generally attempt to identify particular structures (i.e., features) in each field, find the best matches of features across fields, and make comparisons between these matched features based on different attributes (e.g., spatial displacement, orientation, size, etc.). Examples of these types of approaches can be found in Ebert and McBride (2000), Davis *et al.* (2006, 2009), Micheas *et al.* (2007), Baldwin and Lakshmivarahan (2003), Ebert and Gallus (2009), and Lack *et al.* (2009).

Some features-based techniques do not all fit as nicely into the above paradigm. The composite method of Nachamkin (2004, 2009), for example, investigates the distributions of forecasted events relative to observed events and vice versa. Marzban and Sangathe (2006a,b) utilize hierarchical cluster analysis that identifies objects at each “scale” (in this case, scale refers to the number of clusters at each iteration of the hierarchy), and various traditional verification statistics can be applied by defining hits, misses and false alarms by proximity of clusters between two fields. Finally, Wernli *et al.* (2008) take a different approach altogether by defining features within a small region (e.g., a river basin), and focus on three summary statistics pertaining to structure, amplitude and location (giving the technique's short name, SAL) without matching features across fields. The technique is most appropriately applied when the precipitation within the domain is of a single type, and provides useful information that can contrast with results from traditional verification approaches.

d. Field Deformation Approaches

Field deformation approaches attempt to spatially manipulate the forecast field to better match the observed field in an optimal way. In each case, the resulting product is a vector field describing these optimal movements. Methods differ primarily by how they deform the forecast field, and how they summarize the resulting vector field describing the deformations. Field deformation methods were introduced relatively early on (e.g., Hoffman *et al.*, 1995; Alexander *et al.*, 1999). For many of these methods, a finite set of points must first be identified, which can go to the extreme of identifying specific features in the two fields, rendering the technique very similar to a features-based approach. Gilleland *et al.* (2009b) simply use a relatively sparse regular grid of points, and found this approach to be adequate for assessing forecast performance. Optical flow techniques (e.g., Keil and Craig, 2007, 2009; Marzban *et al.*, 2009) do not require the identification of control points at all, utilizing a hierarchical stepping algorithm that makes movements at progressively finer scales. Venugopal *et al.* (2005) introduce a summary measure called the Forecast Quality Index (FQI) that differs considerably from those described above, but we classify it as a field deformation approach because it measures the spatial displacement for the entire field.

3. Comparison of types of methods

Each type of method excels at providing particular information about forecast performance. Methods may give detailed accounts of some types of error while other methods may merely be sensitive to those errors. In other cases, certain types of information are not accounted for or explicitly given by some methods. Gilleland *et al.* (2009) considered a large variety of questions concerning characteristics of the methods; here we focus on how the different types of methods address forecast performance as a function of scale.

As mentioned above, there are different interpretations of the term "scale," and many of the methods explicitly provide information about some definition of scale. Features-based approaches can address individual features of differing scales; scale-separation approaches isolate different scales of features (provided that the chosen spectral scales represent physical features), but do not generally address them individually (the wavelet decomposition approach of Briggs and Levine (1997) has the potential for doing this, but to the best of our knowledge, this is not yet proposed). Features-based methods attempt to identify and compare specific structures between the fields, so that if the structures are physically meaningful, then these methods address performance at different physical scales.

Neighborhood methods are generally applied to increasingly larger neighborhoods of grid points. The information about forecast performance, and indeed scale, depends on the type of smoothing filter. For example, upscaling (e.g., Zepeda-Arce *et al*., 2000; Weygandt *et al*., 2004) gives information about whether the forecast resembles the "truth" field when averaged to coarser scales, whereas the FSS informs about how well the forecast event frequency matches the observed event frequency (because there are numerous neighborhood approaches, we refer the reader to Ebert, 2008 and 2009 for a detailed comparison of these techniques and more references).

The field deformation approaches, in contrast to features-based methods, address the entire field at once, rather than specific structures within the field. While optimization is, therefore, influenced by the entire field, deformations for specific structures may still be obtained. However, to provide useful information about such structures, a human analyst is required. In this sense, the techniques could be used as a way of objectifying subjective evaluations.

Table 1 shows a summary of each category's ability to inform about the specific error types of the geometric ICP cases. All of the methods diagnosed the bias error, and only the field deformation approaches were able to diagnose the aspect ratio error. To be fair, the aspect ratio error in the geometric case was contrived in a way that it could easily be interpreted as rotation error. The features-based methods were not too far off by diagnosing the error as rotation error. Naturally, the displacement errors were discerned well by the displacement methods. The neighborhood and scale-separation approaches do not inform about this type of error, but were sensitive to it.

4. Discussion and Future Directions of the ICP

Numerous methods have been proposed for better evaluating forecasts for an entire field. The ICP was started to help reduce the learning curve for potential users, and to better understand what types of information each method gives; as well as pros and cons of the methods. Initial attention for the ICP has been given to QPF test cases, as well as very simple geometric cases. Actual forecast cases are complemented by realistic cases with known errors. Future work will apply the new verification methods to different types of fields to assess how the methods measure errors in forecasts for other types of variables such as wind and clouds.

All of the new techniques provide useful information about forecast performance when faced with a complete grid of forecasts with an accompanying verification field on the same grid. In particular, bias is handled by all of the methods. Errors of different types may also vary depending on the scale of interest. Neighborhood and scale separation methods are especially well suited for addressing this issue, though there are potentially ways for other types of methods to also account for this behavior. Several of the methods can directly diagnose spatial displacement errors, and many others are sensitive to such errors. In particular, the methods grouped together as features-based and field deformation are designed for this purpose, though they can also address other verification questions in the spatial setting that cannot be addressed by traditional verification methods. For example, the features-based methods can inform about how well a forecast is able to capture various attributes of specific structures within a field (e.g., large storm systems), which may have physical interpretations.

There is much overlap in the information provided by the different techniques. In part, the general categories described in Gilleland *et al.* (2009a) are useful for organizing the numerous methods, but they also give an indication of which method or combinations of methods might be useful to provide a complete picture of forecast performance for a particular application. For example, in planning flight routes, the exact spatial placement of a forecast may be of less interest than the average intensity and spatial extent of a large storm system. In such a case, a displacement method might be useful in discerning how well a forecast performs in spite of such errors. Filter approaches might also be used for this purpose. For example, a scale separation method might be utilized to discern whether a forecast accurately captures the behavior of large-scale weather phenomena. In general, if interest rests on knowing the scale at which a forecast achieves a desired level of skill, then a neighborhood method will be well suited. If it is desired to know how well a forecast captures specific physical scales, then a scale separation or features-based method are recommended. Finally, when interest is in knowing the total amount of spatial displacement error, with amplitude error adjusted for displacement error, then the field deformation methods are ideal.

Acknowledgments.

The authors and participants would like to thank Mike Baldwin for supplying WRF output cases for the ICP. We would also like to thank all of the participants of the inter-

comparison project who took part in the initial planning meeting on 20 February 2007, and the workshops on 14-15 April 2008 and 24-25 August 2009.

References

Ahijevych, D., E. Gilleland, B. Brown, and E. Ebert, 2009: Application of Spatial Forecast Verification Methods to Gridded Precipitation Forecasts. Accepted to *Wea. Forecasting*.

Alexander, G.D., J.A. Weinman, V.M. Karyampudi, W.S. Olson, and A.C.L. Lee, 1999: The effect of assimilating rain rates derived from satellites and lightning on forecasts of the 1993 superstorm. *Mon. Wea. Rev*., **127**, 1433-1457.

Baldwin, M.E. and S. Lakshmivarahan, 2003: Development of an events-oriented verification system using data mining and image processing algorithms. *AMS 2003 Annual Meeting 3rd Conf. Artificial Intelligence*, **4.6**.
URL: http://ams.confex.com/ams/pdfpapers/57821.pdf

Baldwin, M.E. and J.S. Kain, 2006: Sensitivity of several performance measures to displacement error, bias, and event frequency. *Wea. Forecasting*, **21**, 636-648.

Briggs and Levine, 1997: Wavelets and field forecast verification. *Mon. Wea. Rev*. **125**, 1329–1341.

Casati, B., 2009: New Developments of the Intensity-Scale Technique within the Spatial Verification Methods Inter-Comparison Project. Accepted to *Wea. Forecasting*.

Casati, B., G. Ross, and D.B. Stephenson, 2004: A new intensity-scale approach for the verification of spatial precipitation forecasts. *Meteorol. Appl.,* **11**, 141–154.

Davis, C.A., B.G. Brown, and R.G. Bullock, 2006: Object-based verification of precipitation forecasts, Part I: Methodology and application to mesoscale rain areas. *Mon. Wea.Rev*. **134**, 1772–1784.

Davis, C.A., B.G. Brown, R.G. Bullock and J. Halley Gotway, 2009: The Method for Object-based Diagnostic Evaluation (MODE) Applied to WRF Forecasts from the 2005 SPC Spring Program. Accepted to *Wea. Forecasting*.

Ebert, E.E., 2008: Fuzzy verification of high resolution gridded forecasts: A review and proposed framework. *Meteorol. Appl*., **15**, 51-64.

_________, 2009: Fuzzy (neighborhood) verification a strategy for rewarding close forecasts. Accepted to *Wea. Forecasting*.

_________ and W.A. Gallus, Jr., 2009: Toward better understanding of the contiguous rain area (CRA) method for spatial forecast verification. Accepted to *Wea. Forecasting*.

_________ and J.L. McBride, 2000: Verification of precipitation in weather systems: determination of systematic errors. *J. Hydrology*, **239**, 179–202.

Gilleland E., D. Ahijevych, B.G. Brown, B. Casati, and E.E. Ebert, 2009a: Spatial Forecast Verification Methods Inter-comparison. *Wea. Forecasting*, accepted.

Gilleland, E, J. Lindström and F. Lindgren, 2009b: Analysing the image warp forecast verification method on precipitation fields from the ICP. Submitted to *Wea. Forecasting*.

Harris, D., E. Foufoula-Georgiou, K.K. Droegemeier, J.J. Levit, 2001: Multiscale statistical properties of a high-resolution precipitation forecast. *J. Hydrometeorology*, **2**, 406–418.

Hoffman, R.N., Z. Liu, J.-F. Louis, and C. Grassotti, 1995: Distortion representation of forecast errors. *Mon. Wea. Rev.*, **123**:2758—2770.

Jolliffe, I.T. and D.B. Stephenson, 2003: Forecast Verification. A Practitioner's Guide in Atmospheric Science. Wiley and Sons Ltd, 240 pp.

Keil, C., and G.C. Craig, 2007: A Displacement-Based Error Measure Applied in a Regional Ensemble Forecasting System. *Mon. Wea. Rev.*, **135**, 3248–3259.

Keil, C. and G.C. Craig, 2009: A displacement and amplitude error based score employing an optical flow technique. Accepted to *Wea. Forecasting*.

Lack, S., G.L. Limpert and N.I. Fox, 2009: An object-oriented multiscale verification scheme. Accepted to *Wea. Forecasting*.

Lakshmanan, V. and J.S. Kain, 2009: A Gaussian Mixture Model Approach to Forecast Verification. Submitted to *Wea. Forecasting.*

Marzban, C., S. Sandgathe, 2006a: Cluster analysis for object-oriented verification of fields: A variation. *Submitted to Mon. Wea. Rev*.

_________, and _________, 2006b: Cluster analysis for verification of precipitation fields. *Wea. Forecasting*, **21**, 824-838.

_________, and _________, 2009: Verification with variograms. Accepted to *Wea. Forecasting*.

Marzban, C., S. Sandgathe, H. Lyons, and N. Lederer, 2009: Three Spatial Verification Techniques: Cluster Analysis, Variogram, and Optical Flow. Accepted to *Wea. Forecasting*.

Mesinger, F., 2008: Bias adjusted precipitation threat scores. Adv. Geosciences **16,** 137--143.

Micheas, A.C., N.I. Fox, S.A. Lack, and C.K. Wikle. 2007: Cell identification and verification of QPF ensembles using shape analysis techniques. *J. of Hydrology*, **343**, 105-116.

Mittermaier, M.P. and N. Roberts, 2009: Intercomparison of Spatial Forecast Verification Methods: Identifying skillful spatial scales using the Fractions Skill Score. Accepted to *Wea. And Forecasting*.

Nachamkin, J.E., 2004: Mesoscale verification using meteorological composites. *Mon. Wea. Rev.*, **132**, 941-955.

Nachamkin, J.E., 2009: Application of the Composite Method to the Spatial Forecast Verification Methods Inter-Comparison Dataset. Accepted to *Wea. Forecasting*.

Roberts NM and HW Lean, 2008. Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. Mon. Wea. Rev., **136**:78—96.

Tustison, B., D. Harris, and E. Foufoula-Georgiou, 2001: Scale issues in verification of precipitation forecasts. *J. Geophysical Research*, **106**, 11775-11784.

Venugopal, V., S. Basu, E. Foufoula-Georgiou, 2005: A new metric for comparing precipitation patterns with an application to ensemble forecasts. *J. Geophys. Res*. **110**, D08111.

Wernli, H., C. Hofmann, and M. Zimmer, 2009: Spatial forecast verification methods inter-comparison project application of the SAL technique. Accepted to *Wea. Forecasting*.

_________, M. Paulat, M. Hagen and C. Frei, 2008: SAL - A novel quality measure for the verification of quantitative precipitation forecasts. *Mon. Wea. Rev*., **136**, 4470-4487.

Weygandt, S.S., A.F. Loughe, S.G.Benjamin, and J.L. Mahoney, 2004: Scale sensitivities in model precipitation skill scores during IHOP. *22nd Conference on Severe Local Storms*. American Meteorological Society: Hyannis, Massachusettes, 4-8 October 2004.

Wilks, D.S., 1995 1995: Statistical Methods in the Atmospheric Sciences. An Introduction. Academic Press, San Diego, 627 pp.

Zepeda-Arce, J., E. Foufoula-Georgiou, and K.K. Droegemeier, 2000: Space-time rainfall organization and its role in validating quantitative precipitation forecasts. *J. Geophysical Res.*, **105**(D8):10,129-10,146.

List of Tables

Table 1: Summary of general results by category for the types of errors represented by the geometric ICP test cases. Types (*) indicates that the method type may be sensitive to the type of error, but does not directly measure it.

	Neighborhood methods	**Scale-separation methods**	**Features-based methods**	**Field deformation methods**
Bias	Yes	Yes	Yes	Yes
Displacement	No*	No*	Yes	Yes
Aspect ratio	No	No	No*	Yes

List of Figures

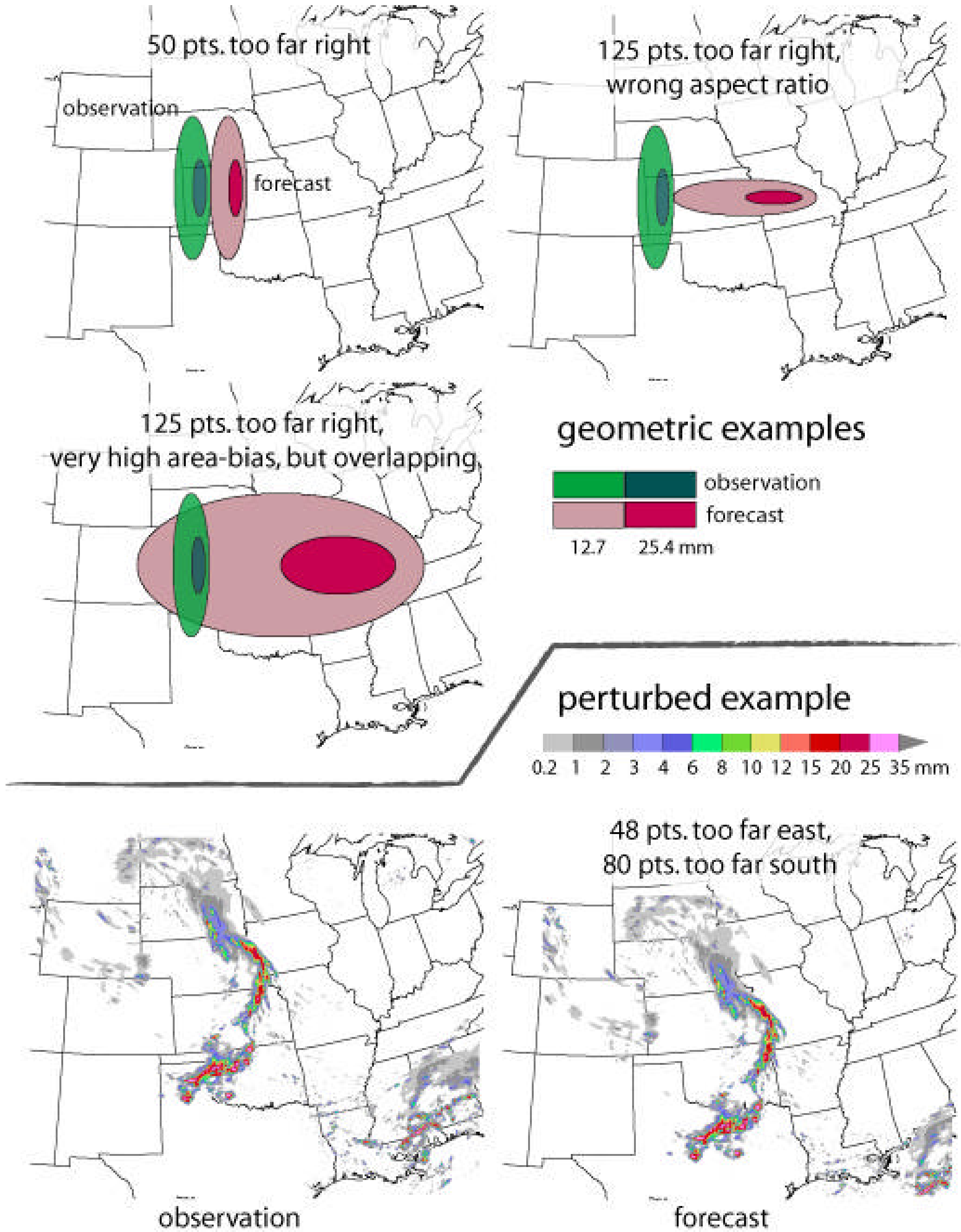

FIG. 1. Examples of the artificial cases used to test the various verification methods. The top three panels show examples of the geometric cases, and the bottom two panels show an example from the set of perturbed cases with the observation on the left and the perturbed forecast on the right.

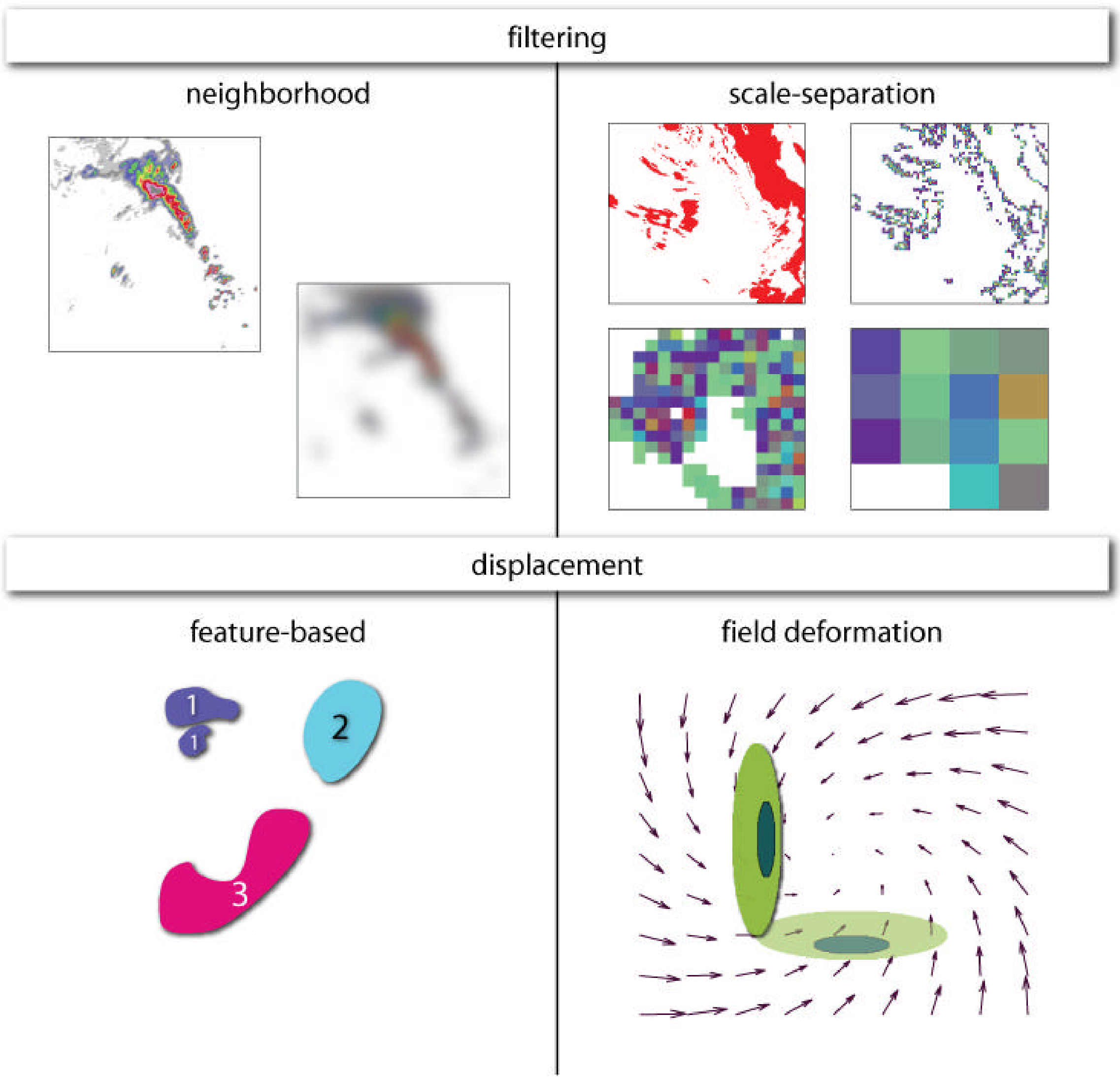

FIG. 2. A schematic showing the differences between the various types of methods (reproduced from Gilleland *et al*., 2009a).