VERIFYING FORECASTS SPATIALLY

by Eric Gilleland, David A. Ahijevych, Barbara G. Brown, and Elizabeth E. Ebert

An international project tests new spatial forecast verification methods to find out how they handle different types of forecast error and what they tell us about forecast performance.

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., the probability of detection (POD) and the Gilbert skill score (GSS)] and diagnostics (e.g., the relative operating characteristic) of forecast performance (for more on traditional verification, see, e.g., Wilks 1995; Jolliffe and Stephenson 2003). 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 overpredict 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 used regularly at meteorological centers around the world. Because the methods are relatively new, however, it is not always clear which 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 about which a user should know? Such questions provided the impetus for the Spatial Forecast Verification Methods Inter-Comparison project (ICP; details at 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.

TEST CASES. 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 (QPFs) are being used, with plans to incorporate more varied cases in the future. The cases
under study now include 24-h forecasts of 1-h precipitation from three configurations of the Weather Research and 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 (Fig. 1a) are simple elliptically shaped precipitation patterns with common forecast errors and provide useful information about the output of each method. Perturbed “real” cases (Fig. 1b) illustrate the capabilities of each method with more complex precipitation scenes and similar errors to the geometric cases. For nine real WRF cases, a subjective evaluation was carried out to compare the human ranking to the spatial verification methods, though 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).

THE METHODS. Most of the techniques proposed can be classified into one of the following four categories: a) scale separation (or decomposition), b) neighborhood (or fuzzy), c) features based (or objects based), and d) field deformation. The first two could be further generalized as filter methods where the scale-separation methods take advantage of bandpass 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. For the filter methods, the summary statistics are applied at different scales; in the case of the neighborhood methods (Fig. 2a), these statistics are calculated for a range of neighborhood sizes, whereas for scale separation (Fig. 2b) they are calculated for different spectral bands that isolate phenomena of a particular size. To illustrate the scale-separation approach, Fig. 2b shows the binary difference between a forecast and observation, and constituent wavelet components. The displacement methods attempt to

**Fig. 1.** Select examples of the artificial cases used to test the various verification methods. (a) Three geometric cases and (b) one of the perturbed cases are shown. The full set of test cases is described in Ahijevych et al. (2009).
displace the forecast field spatially to better match the observations. Information about the amount of displacement that is necessary can be gleaned along with various other diagnostics and summary statistics. The primary difference between the methods is that the features-based approaches (Fig. 2c) identify individual features (or objects) within a field and analyze these structures separately, whereas the field-deformation methods (Fig. 2d) apply to the entire field as a whole. The following subsections provide brief descriptions of the methods by type. A more complete literature review and qualitative comparison of the methods can be found in Gilleland et al. (2009).

**Neighborhood approaches.** Neighborhood approaches differ from one another primarily by the type of smoothing filter applied. The filter is applied to the forecast field, and in most cases also the observed field (some neighborhood methods verify forecasts against point observations), and the summary statistics (e.g., traditional verification statistics) are applied to the filtered fields. Further, most filters preserve the peak values, which are important for capturing forecasting capabilities for extreme events such as severe winds or large hail. 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 (2010) 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 being able to identify which scales have useful skill.

**Scale-separation approaches.** Scale-separation techniques are not new to forecast verification (Briggs and Levine 1997; Tustison et al. 2001). Typically, each field is decomposed using some type of bandpass filter (Fourier, wavelets, etc.), and the two fields are compared using traditional verification techniques at each spectral scale. Note that the term “scale” used here can be linked to physical features, such as 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 investigate a forecast’s ability to reproduce the observed variability-scale structure of the observed field. The methods of Harris et al. (2001) and Casati et al. (2004) are examples of scale-separation techniques.

Lack et al. (2010) 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.

**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 (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. (2010).

Some features-based techniques do not 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 (2006, 2008) 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 the proximity of the clusters between the two fields. Finally, Wernli et al. (2008, 2009) take a different approach altogether by defining features within a small region (e.g., a river basin), and focusing on three summary statistics pertaining to structure, amplitude, and location (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 complement results from traditional verification approaches.

**Field-deformation approaches.** Field-deformation methods were introduced for meteorological verification by Hoffman et al. (1995) and Alexander et al. (1999). These 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. 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 to be very similar to that of a features-based approach. Gilleland et al. (2010) 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) 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), which 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.

**COMPARISON OF TYPES OF METHODS.** Each type of method excels at providing particular information about forecast performance. Some 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 a representative from each category is demonstrated on an example case.

Each verification method has a different look and feel. To illustrate this with a concrete example, one representative method from each of the four broad verification categories is applied to the same forecast (Fig. 3). The example comes from one of the nine real test cases described in Ahijevych et al. (2009). At the top of the figure, the 1-h accumulated precipitation fields are shown for the 24-h WRF National Centers for Environmental Prediction (NCEP) forecast and stage II observation. The GSS (equivalent to the equitable threat score) and frequency bias (ratio of forecast area to observed area) are calculated for a precipitation threshold of 5 mm. To illustrate the neighborhood approach (Ebert 2009), the precipitation fields are filtered with a 132 km × 132 km moving average window (top left) and a new GSS and bias are calculated. The GSS increases from 0.06 to 0.17, and the bias improves from 1.87 to 1.14. This relation between skill score and neighborhood size is typical.
for high-resolution precipitation forecasts; as one averages out the small-scale variability associated with grid-scale precipitation, the skill scores improve. The models may not pinpoint the exact location of each storm cell, but they can correctly place the overall envelope of precipitation. The neighborhood approach filters out the high-frequency variability and can allow a high-resolution model to be compared to a low-resolution model without being penalized for small errors.

The scale-separation approach is similar to the neighborhood approach, but it uses a Fourier or wavelet transform to decompose the precipitation fields. The 512- and 64-km Haar wavelet components of the binary difference field are shown in Fig. 3c. As in Casati (2010), the forecast and observation fields are first converted to binary (0/1) fields, and then the difference field is decomposed. A greater proportion of the difference field is associated with the 64-km component (23%) than with the 512-km component (3%), suggesting the model forecast has more error at 64-km scales than at 512 km. Additional statistics are available for the remaining wavelet components, but they are not shown.

In Fig. 3d, the object-based method isolates the main features and then matches the resultant objects. In this example from the method for object-based diagnostic evaluation (MODE; Davis et al. 2009), the red forecast object is matched with the red object in the observed field. Object attributes such as centroid displacement, orientation angle difference, and median intensity are summarized below the two fields. Objects that are not matched are dark blue. In the forecast field, these unmatched objects can be considered false alarms, and in the observed field, they can be considered misses. In this case, there are numerous false alarms in the southeastern United States and one miss in Texas.

The final category of verification method illustrated in Fig. 3 is field deformation. In Fig. 3e, an optical flow technique is applied to the forecast field to make it look more like the observed field (adapted from Keil and Craig 2009). In the full displacement and amplitude score (DAS) calculation described in Keil and Craig (2009), the observation is also warped to look like the forecast (not shown). Some of the statistics that come out of the DAS method\(^1\) are mean displacement error (104.4 km) and root-mean-square amplitude error of the morphed fields (6.79 mm). The field-deformation methods diagnose similar forecast errors to the object-based methods, but they treat the field as a whole instead of separate objects.

Table 1 summarizes the methods by category for several issues relevant to forecast quality. It should be noted that while the table gives yes or no answers for categories of methods, the answer may not apply to a particular method in the category. For example, while the displacement methods are not generally set up for informing about scales where a forecast has skill, it is certainly feasible to apply any of the methods to particular scales (an upscaled field, a detailed field from a wavelet decomposition, etc.). The method by which MODE, a features-based displacement method, creates objects involves a convolution of the field using a user-defined radius. The amount of convolution can be considered as a scale (similar to upscaling) so that skill at different scales can be obtained.

All of the methods supply information about intensity errors, but it should be noted that they do so in very different ways. A deformation approach, for example, typically computes traditional verification scores that pertain to intensity for pre- and postdeformed fields. Neighborhood methods also typically provide traditional verification scores, but on filtered fields of different scales. In this way, information is obtained about the scales where a forecast has skill.

---

**Table 1. Summary of the types of information that the general categories of methods can provide.** Entries should be interpreted as being a general characteristic of the method type, and it should be understood that there are exceptions among individual methods; see the text for more information.

<table>
<thead>
<tr>
<th>Category</th>
<th>Scales with skill</th>
<th>Location errors</th>
<th>Intensity errors</th>
<th>Structure errors</th>
<th>Occurrence (hits, misses, and false alarms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neighborhood</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Scale separation</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Features based</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Deformation</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

\(^1\) In Keil and Craig (2009), the displacement error is then normalized by the maximum search distance \((D_{\text{max}} = 360 \text{ km})\) and the amplitude error is normalized by the characteristic intensity \((I_{\text{c}} = 6.23 \text{ mm})\).
Fig. 3. (a) Each of the four verification categories illustrated in Fig. 2 is applied to the same test case. (b) Illustrated is a neighborhood approach, in which the forecast and observation fields are upscaled with a 132 km × 132 km averaging filter. (c) The scale-separation approach is illustrated by showing two components of the wavelet decomposition. (d) The object-based method matches the red forecast and observed object. (e) The field-deformation example uses optical flow (Keil and Craig 2009) to morph the forecast to the observation.
Contingency table information (i.e., hits, misses, false alarms, and correct negatives) and binary scores are naturally summarized by neighborhood and scale-separation methods. Features-based methods provide a new way of determining such errors, where matched objects can be thought of as hits and unmatched objects as misses or false alarms.

**DISCUSSION AND FUTURE DIRECTIONS OF THE ICP.** 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 the 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., orientation of 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. (2009) are useful for organizing the numerous methods, but they also give an indication of which method or combinations of methods might be useful for providing 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 is recommended. Finally, when the 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.

Many of the newly proposed techniques are gaining momentum, and some are already in operational use [e.g., the Met Office has been using the fractional skill score approach since 2007 (M. P. Mittermaier 2010, personal communication); SAL is used for internal operational verification at the Finnish Meteorological Institute (M. Zimmer 2010, personal communication); MODE is in operational use in China and is being considered for operational use in the United States; and contiguous rain area (CRA) is used for precipitation verification at the Bureau of Meteorology]. Other methods are beginning to be applied in nonoperational settings [e.g., SAL is being studied by numerous scientists and is included as part of the High-Resolution Limited-Area Model (HIRLAM) community internal verification toolbox; MODE is being used in various research projects and in demonstrations for the U.S. Hydrometeorology and Hazardous Weather Testbeds (e.g., see http://verif.rap.ucar.edu/eval/hmt/2010/graphics/); the Met Office has used the intensity-scale method for nonoperational purposes, and the CRA is being implemented for assessing dust forecasts (M. P. Mittermaier 2010, personal communication). In addition to its operational use, the FSS is being used in evaluation of hazardous weather forecasts in the United States (e.g., Schwartz et al. 2010)].

Future spatial verification methods are likely to combine filter and displacement approaches. While it is possible to apply one of each method, using them in concert can potentially provide a more complete picture of forecast performance. For example, one might want to know how serious location errors are for different scales. Does a forecast predict local phenomena well after accounting for a larger-scale displacement? It is certainly natural to apply any displacement method to scales obtained through filtering or multiresolution analyses. Such hybrid
methods are already being proposed (e.g., Lack et al. 2010).

Software for some of the techniques is available through the ICP Web site. The intensity-scale technique is available as part of the R (R Development Core Team 2009) package verification, as well as in the Model Evaluation Tools verification software (www.dtcenter.org/met/users/), which also includes software for employing MODE.

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 intercomparison 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. Figures 2 and 3 were derived from output from the MET verification software package (www.dtcenter.org/met/users), which was developed with the support of the U.S. Air Force Weather Agency and the National Oceanic and Atmospheric Administration.

REFERENCES


Micheas, A. C., N. I. Fox, S. A. Lack, and C. K. Wikle, 2007: Cell identification and verification of QPF.


AMS MEMBERS
GIVE A GREAT GIFT AT A GREAT PRICE

Looking for the perfect present for the weather enthusiast in your life? Want to make a valuable contribution to your local library or community college? Send a subscription to Weatherwise magazine (6 issues) for just $24.95*—That’s nearly 50% off the list price!

Written for a general audience, Weatherwise offers a colorful and nontechnical look at recent discoveries in meteorology and climatology. Check out the latest table of contents at www.weatherwise.org.

Want your own? Then order a personal subscription at the same great price.

Use your 2011 dues renewal to place all of your Weatherwise orders today!

*Cost for delivery outside of the U.S. is $40.95. Weatherwise is available to AMS Members through a cooperative agreement with Taylor & Francis Group LLC, the publishers of Weatherwise.
GLOSSARY OF METEOROLOGY

Second Edition

In 1959, the AMS published the *Glossary of Meteorology*. Containing 7900 terms, more than 10,000 copies have been sold over four decades through five printings. It is a tribute to the editors of the first edition that it has withstood the test of time and continued to be among the leading reference sources in meteorology and related sciences.

Now, over five years in the making, the second edition is available. The volume contains over 12,000 terms, including those from “new” disciplines, such as satellite meteorology and numerical weather prediction. In addition, related oceanographic and hydrologic terms are defined.

The *Glossary of Meteorology*, Second Edition, was produced by an editorial board comprised of 41 distinguished scientists and the participation of over 300 contributors. The CD-ROM version is compatible with Windows, Macintosh, and most UNIX platforms, and features hyperlinked cross-references.


HARDBOUND BOOK
$85 List
$60 AMS Members
$35 AMS Student Members

CD-ROM
$95 List
$65 AMS Members
$35 AMS Student Members

Order online: www.ametsoc.org/amsbookstore
or see the order form at the back of this issue.
You have arrived at a page where you face a choice. As a reader, you are accustomed to choices. The choice on the following pages, however, is bigger than most choices in this publication. At the very least your decision affects the coherence of what you are about to read. So a word of explanation is in order.

The next four papers deal with the future of prediction in your field, but the relationship between the papers may not be clear if you pick and choose among them. The first paper (Shapiro et al.) is an overarching vision for improving Earth-system prediction. You might find yourself daunted by this challenging perspective, which calls for the Earth-system-wide integration described in the following paper (Nobre et al.). Maybe you will gravitate instead toward the more familiar territory of subseasonal and seasonal forecasting (Brunet et al.). Or you’ll slide past the context to get to the proposal for a leap in the capabilities of modeling facilities (Shukla et al.).

Clearly, you can choose amongst these papers, but that is not what the authors had in mind. These articles were submitted and considered as if they were one, each depending on the success of the other for the strength of their common message. No one of these articles would have been accepted without the other three.

Actually, maybe that’s the whole point of this choice. The authors are presenting their proposal over the course of four articles, instead of one, because of the enormity of the implications for the entire prediction enterprise. The articles expose the distinct parts that must be fused to make the proposal work. That ultimately requires more space for a single, multifaceted idea than is the norm for BAMS. And that leaves it to you, the reader, to choose whether or not to do the integrating. Your choice is central to making the initiative succeed, even if only on paper this time.

It is also likely the Earth-system prediction concept posed to your community in these pages will require not just a one-time choice now but also a lasting resolve to integrate parts into a new whole. Thus you will need to ask yourself, can progress in science be enhanced by creating a new context that intensifies synergy of interests, disciplines, expertise, scales, methodologies, and resources? Does this cohesive context lead logically to a new way of providing services? Does it produce a better way of asking questions and exploring unknowns?

The authors have good reason for their faith in your desire to integrate. The meteorological and oceanographic communities have consistently shown that they have a knack for close collaboration, which is an essential facet of integration. Since 1780, with the formation of the Societas Meteorologica Palatina, the science of weather has depended upon international cooperation for data collection. Matthew Maury is lionized for similarly establishing the collaborative context in oceanography. Now you are being asked to build on this heritage of international cooperation and expand it as an intellectual cooperation to entwine related fields of inquiry with a common purpose.

If 40 authors from diverse countries, institutions, and research agencies can unite to present such a vision, as they do in the following pages, then maybe their proposal stands a chance to succeed. These authors are now asking you, the reader, to take the next step into the future by choosing to read four papers as one. Authors rarely ask so much, but, then again, scientific communities rarely face such choices.

Jeff Rosenfeld
Editor-in-Chief