New Developments of the Intensity-Scale Technique within the Spatial Verification Methods Intercomparison Project

B. CASATI

Consortium Ouranos, Montreal, Quebec, Canada

(Manuscript received 6 January 2009, in final form 3 August 2009)

ABSTRACT

The intensity-scale verification technique introduced in 2004 by Casati, Ross, and Stephenson is revisited and improved. Recalibration is no longer performed, and the intensity-scale skill score for biased forecasts is evaluated. Energy and its percentages are introduced in order to assess the bias on different scales and to characterize the overall scale structure of the precipitation fields. Aggregation of the intensity-scale statistics for multiple cases is performed, and confidence intervals are provided by bootstrapping. Four different approaches for addressing the dyadic domain constraints are illustrated and critically compared.

The intensity-scale verification is applied to the case studies of the Intercomparison of Spatial Forecast Verification Methods Project. The geometric and synthetically perturbed cases show that the intensity-scale verification statistics are sensitive to displacement and bias errors. The intensity-scale skill score assesses the skill for different precipitation intensities and on different spatial scales, separately. The spatial scales of the error are attributed to both the size of the features and their displacement. The energy percentages allow one to objectively analyze the scale structure of the fields and to understand the intensity-scale relationship. Aggregated statistics for the Storm Prediction Center/National Severe Storms Laboratory (SPC/NSSL) 2005 Spring Program case studies show no significant differences among the models' skill; however, the 4-km simulations of the NCEP version of the Weather Research and Forecast model (WRF4 NCEP) overforecast to a greater extent than the 2- and 4-km simulations of the NCAR version of the WRF (WRF2 and WRF4 NCAR). For the aggregated multiple cases, the different approaches addressing the dyadic domain constraints lead to similar results. On the other hand, for a single case, tiling provides the most robust and reliable approach, since it smoothes the effects of the discrete wavelet support and does not alter the original precipitation fields.

1. Introduction

Progress in numerical weather prediction (NWP) and the advent of high-resolution precipitation forecasts have raised the need for more informative verification approaches capable of dealing with the complex spatial structure of such fields. Several new spatial verification approaches have been developed in the past decade to account for the following: the space–time uncertainties related to small displacements between features in the observation and forecast fields (e.g., Ebert 2008, and references therein); the scale structure of the fields and scale dependency of predictability and skill (e.g., Harris et al. 2001; Casati et al. 2004); the existence of features with possible displacement, extent, and intensity errors (e.g., Ebert and McBride 2000; Davis et al. 2006); and the intrinsic coherent spatial structure of the forecast fields, which can be deformed to match the observation field (e.g., Hoffman et al. 1995; Keil and Craig 2007). These new spatial verification methods have been grouped into four categories: *neighborhood* (or *fuzzy*), *scale-separation* (or *scale decomposition*), *feature-based* (or *object oriented*), and *field-deformation* approaches (Casati et al. 2008; Gilleland et al. 2009). The Intercomparison of Spatial Forecast Verification Methods (Gilleland et al. 2009) is a metaverification project that aims to analyze the performance of these new techniques on a dataset of common case studies (Ahijevych et al. 2009). The goal is to better understand the capabilities and information provided by each method, in order to guide users with specific applications to the choice of the most adequate verification approach. This article proposes some improvements to the intensity-scale (IS) verification technique introduced by Casati et al. (2004) within the

Corresponding author address: Dr. B. Casati, Consortium Ouranos, 550 Sherbrooke West, 19th Fl., Montreal QC H3A 1B9, Canada.
E-mail: b.casati@gmail.com

DOI: 10.1175/2009WAF2222257.1

context of the Intercomparison of Spatial Forecast Verification Methods.

The IS technique belongs to the group of scale-separation verification methods (Casati et al. 2008; Gilleland et al. 2009). These methods decompose forecast and observation fields into scale components by using a spatial filter (e.g., wavelets, Fourier transforms). Verification with traditional scores is then performed for each individual scale (e.g., Briggs and Levine 1997; Casati et al. 2004). Scale-separation approaches can assess the scale dependency of the error and the no-skill to skill transition scale. Moreover, Zepeda-Arce et al. (2000) and Harris et al. (2001) analyze the spatial structure of the forecast and the observation fields by assessing some scale-invariant parameters related to the spatiotemporal organization of the precipitation fields. Note that scale-separation approaches rely on a single-band spatial-scale filter, whereas neighborhood verification methods [Ebert (2008, 2009); see, e.g., the fraction skill score introduced by Roberts and Lean (2008)] operate with a low-bandpass filter (i.e., smoothing). Therefore, scale-separation approaches are capable of isolating individual wavelengths, which are associated with weather phenomena of different scales (e.g., fronts or convective cells), and can provide information on the forecast errors and skill for individual scales, separately. Neighborhood verification methods, on the other hand, do not aim to separate the scales, but assess the critical smoothing scale (or resolution) above which the desired skill is achieved. Scale-separation and neighborhood verification approaches differ fundamentally because of their different definitions of ''scale,'' which lead to different interpretations of the verification results. More discussion on these differences can be found in Gilleland et al. (2009), Ebert (2009), and Casati et al. (2008).

The IS technique assesses the forecast skill for different spatial scales and precipitation intensities. The scales are obtained by a 2D Haar wavelet filter applied to thresholded forecast and observation fields. The thresholding process allows the IS technique to bridge categorical and scale-separation verification approaches. Haar wavelets and the categorical approach were chosen in the design of the IS verification technique because they are robust and well suited for analyzing sparse precipitation fields characterized by spatial discontinuities and highly skewed intensity distribution.

The IS technique is described here from a new perspective, which aims to be complementary to that in Casati et al. (2004). In particular, a more standard wavelet approach is used to define the IS scale components. Moreover, new developments in the IS verification method are proposed in order to address issues raised in recent studies that made use of the IS technique (e.g., Mittermaier 2006; Cisma and Ghelli 2008). One change is that the bias is no longer removed from the forecast prior to skill assessment. The energy is instead introduced to assess the bias on different scales and for different thresholds. The scale structure is also evaluated by the energy percentages. A method for aggregating IS verification statistics for multiple case studies is presented to respond to operational verification needs. Finally, the constraints dictated by the need of a dyadic domain for discrete wavelet transforms are addressed with different approaches.

The IS technique is reviewed and its new developments are illustrated in section 2. The IS verification is then applied to the spatial verification methods intercomparison case study dataset. A detailed description of these cases can be found in Ahijevych et al. (2009). The IS verification results for the geometric cases, the synthetically perturbed case, and the Storm Prediction Center/National Severe Storms Laboratory (SPC/NSSL) 2005 Spring Program case studies are presented in section 3. Discussion and conclusions are given in section 4.

2. The method

The following description of the IS technique intends to be complementary to that in Casati et al. (2004). In particular, the definition of the scale components and the wavelet filter are treated differently. In the original paper, the scale components were obtained as the difference of fields smoothed at the resolutions of 2^j and 2^{j-1} grid points. In this work, the scale components are obtained by inverting the 2D Haar discrete wavelet transform for the vertical, horizontal, and diagonal wavelet coefficients and then summing the fields reconstructed for the three wavelet orientations, for each individual scale. Note that the scale components obtained with the two procedures are identical. Despite the similarity between the wavelet filter as described in Casati et al. (2004) and Laplacian pyramids (Burt and Adelson 1983), the IS technique scale components differ substantially from the ones obtained by a Laplacian pyramid. In fact, wavelet-based scale components (such as the IS technique scale components) are orthogonal, whereas Laplacian pyramid components are not (Mallat 1989). Note that orthogonality is a key feature of the IS technique, since this enables the IS statistics to be additive [e.g., Eq. (1)].

Casati et al. (2004) applied the IS verification to preprocessed and recalibrated (unbiased) data. The aim of the preprocessing was to normalize the data and define thresholds so that each categorical bin had a similar sample size, whereas the recalibration was performed to eliminate the marginal distribution bias. Preprocessing and recalibration are not strictly necessary for the IS

technique. In this study, neither preprocessing nor recalibration is performed, and the IS approach is applied for empirically chosen categorical thresholds to biased forecasts. Note that the IS skill score for biased forecasts differs from that for unbiased forecasts for an extra scale component: in the following description, particular attention is given to this additional bias scale component.

a. The intensity-scale skill score

For each threshold, the forecast and observation fields are transformed into binary fields: where the gridpoint precipitation value exceeds the threshold, it is assigned 1; where the threshold is not exceeded, it is assigned 0. Figures 1a and 1b illustrate examples of a forecast and the observation fields and Figs. 1c and 1d show their corresponding binary fields for a threshold of 1 mm h^{-1}. This case is one of the NIMROD case studies analyzed in Casati et al. (2004). This case was also used by Ebert (2008) to illustrate different neighborhood verification methods and is used in this section to illustrate the IS approach. A 3-h lead-time forecast of precipitation rate produced by NIMROD (Golding 2000) is verified against its corresponding radar-based analysis on a 5-km-resolution spatial domain over the United Kingdom. The interesting feature this case shows is an intense storm of the scale of 160 km, which is displaced almost its entire length. The displacement error is clearly visible from the binary field difference (Fig. 1e) and the contingency table image (Fig. 1f) illustrating the counts of the contingency table (Table 1) obtained for the same threshold.

The binary forecast and observation fields obtained from the thresholding are then decomposed into the sum of the components on different scales. The scale components are obtained as follows. First, a 2D Haar discrete wavelet transform (Mallat 1989; Daubechies 1992) is applied to the field. Wavelet coefficients for the diagonal, horizontal, and vertical 2D wavelets are so obtained. Then, for each individual scale, the inverse discrete wavelet transform is applied to the diagonal, horizontal, and vertical wavelet coefficients. Three reconstructed fields, corresponding to the three orientations, are so obtained in the data space, for each scale. These three fields are then summed, to obtain the corresponding scale component. Figure 2 shows the wavelet scale components of the binary forecast and the observation for the NIMROD case study. The Haar wavelet filter, as any spectral filter, isolates features of different spatial scales. For the NIMROD case study analyzed, the displaced intense storm is identified by large positive values (the black square) in the scale component corresponding to 160 km because of its size (160 km). Note that its positions in the forecast and the analysis scale component are different, due to the 160-km displacement.

The wavelet transform is a linear operator: this implies that the difference between the spatial scale components of the binary forecast and the observation fields (Fig. 2) is equal to the spatial scale components of the binary field difference (Fig. 3). The IS skill score considers the scale components of the binary field difference, which can be obtained either from the wavelet decomposition of the binary field difference (as in Casati et al. 2004) or from the difference of the scale components of the binary forecast and observation fields (as illustrated here). The scale components of the binary field difference for the NIMROD case study (Fig. 3) exhibit a large error at the scale of 160 km, due to the storm being displaced. The scale structure of the error is affected by both the displacement of the feature (160 km) and by its size (160 km).

The IS scale components are fields with different resolutions. For a field defined over a square domain of $2^L \times 2^L$ grid points, there are $L + 1$ scale components. The first L scale components [referred to as *mother wavelet components* in Casati et al. (2004)] are obtained by applying the inverse discrete wavelet transform to the vertical, horizontal, and diagonal wavelet coefficients, and then summing the reconstructed fields obtained from the three wavelet orientations, for each individual scale. These components are produced by wavelets that have square support with dimensions equal to 2, 4, 8, ..., 2^L grid points, and resolutions equal to 1, 2, 4, ..., 2^{L-1} grid points, respectively. The $(L + 1)$th scale component [referred to as the *father wavelet component* in Casati et al. (2004)] is obtained by reconstructing the scaling function with support equal to $2^L \times 2^L$ grid points (i.e., the whole domain). This scaling function component for the Haar wavelet has a resolution equal to 2^L grid points. Throughout this article, the IS scale components $l = 1, 2, 3, \ldots, L + 1$ are uniquely identified by their resolution of $2^{l-1} = 1, 2, 4, \ldots, 2^L$ grid points.

The Haar scaling function component is a constant field over the $2^L \times 2^L$ gridpoint domain, with a value equal to the field mean. For the binary forecast and observation, these are equal to $r = (a + b)/n$ and $s = (a + c)/n$, that is, the proportion of forecast and observed events above the threshold, as evaluated from the contingency table counts. These relations can be easily shown by evaluating the mean of the 0/1 binary fields, and are intuitive when comparing forecast and observation binary fields to their corresponding contingency table image (Figs. 1c, 1d, and 1f). The difference between the scaling function components of the forecast and observation fields (i.e., Fig. 3, scale 9) provides information on the whole-field bias. For an unbiased forecast, the scaling function components of the forecast and observation are equal, and the scaling function component of the field difference is zero

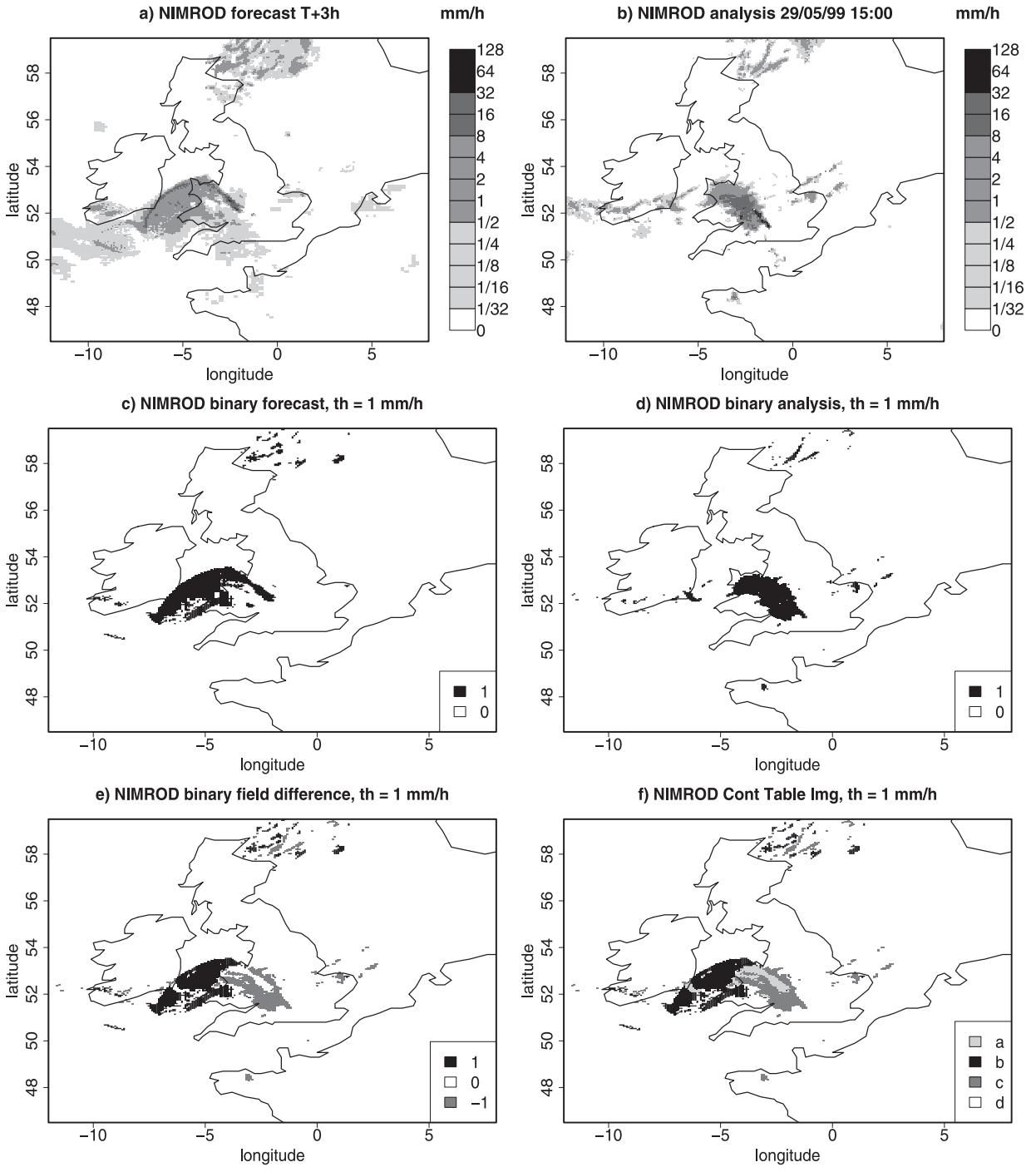

FIG. 1. NIMROD case study: (a) 3-h lead-time forecast and (b) corresponding verifying analysis. Binary fields for the (c) forecast and (d) analysis for a threshold of 1 mm h^{-1}. (e) Binary field difference and (f) corresponding contingency table image.

(no error). Since Casati et al. (2004) considered unbiased forecasts, the (zero) scaling function component of the error was not included in the IS skill score evaluation (it would have been trivially skillful). In the present study, on the other hand, this component is nonzero (since biased forecasts are verified) and is therefore considered in the IS skill score evaluation.

For each threshold and for each scale component, the mean square error (MSE) is evaluated. Note that the MSE can be evaluated either from the scale components of the

TABLE 1. Contingency table: the counts a, b, c, and d correspond to the hits, false alarms, misses, and correct rejections defined for the forecast (F) and observation (O) exceeding the threshold u. The total number of sampled events is equal to $n = a + b + c + d$.

	$O \geq u$	$O < u$	
$F \geq u$	a	b	$a + b$
$F < u$	c	d	$c + d$
	$a + c$	$b + d$	n

binary forecast and observation fields (i.e., the components in Fig. 2), or from the mean of the squared values of the scale components of the binary field difference [i.e., the components in Fig. 3, as in Casati et al. (2004)]. We denote this MSE as $\mathrm{MSE}_{u,l}$ to indicate its dependency on the threshold u and scale l. Figure 4a shows the $\mathrm{MSE}_{u,l}$ for the NIMROD case study. The error is large for small thresholds and decreases as the threshold increases. This behavior is due to the dependence of the error on the frequency of events in the binary fields: the smaller the threshold, the more events will exceed it and, therefore, the larger the error. The threshold dependency of the $\mathrm{MSE}_{u,l}$ can be eliminated by normalization, as explained in the following paragraph.

The scale components in Fig. 2 add up to the original forecast and analysis binary field (Figs. 1c and 1d). Similarly, the scale components in Fig. 3 add up to the original binary field difference (Fig. 1e). Because of the orthogonality of the discrete wavelet transform, the additive properties of the scale components transfer to mean square statistics (see Casati et al. 2004, section 3.2.2). The sum of the MSE of the scale components is therefore equal to the MSE of the original binary fields:

$$\mathrm{MSE}_u = \sum_{l=1}^{L+1} \mathrm{MSE}_{u,l}. \tag{1}$$

This enables one to calculate, for each threshold, the percentage of the total MSE_u that each scale contributes:

$$\mathrm{MSE}\%_{u,l} = \left(\frac{\mathrm{MSE}_{u,l}}{\mathrm{MSE}_u}\right) \times 100. \tag{2}$$

Figure 4b shows the MSE percentages for each threshold and scale for the NIMROD case study. The $\mathrm{MSE}\%_{u,l}$ of precipitation fields usually exhibits small errors on large scales (large scales are more predictable and large displacements seldom occur) and large errors on small scales (usually due to many small-scale displacements), with the largest error associated with the smallest scale and highest thresholds (intense small-scale events are the least predictable). Moreover, the NIMROD case study exhibits a large error at 160 km for thresholds between ½ and 4 mm h^{-1}: such behavior is specific to this case and is due to the 160-km intense storm being displaced almost its entire length.

To define the IS skill score, the MSE for a random binary forecast and the observation fields needs to be estimated. This random binary MSE can be estimated with an approach that is similar to that in section 3.2.3 of Casati et al. (2004), and assuming that random binary forecast and observation fields are independent Bernoulli-distributed variables with expectations equal to r and s, respectively. In the present study, however, an alternative approach is shown, in order to highlight the strong link between categorical and IS verification statistics. Let us recall that the MSE of the original binary fields (which is denoted by MSE_u to indicate its dependency on the threshold u) is equal to the sum of the proportion of misses (c/n) and false alarms (b/n) for the contingency table obtained with the same threshold:

$$\mathrm{MSE}_u = \frac{(b + c)}{n}. \tag{3}$$

This relation follows from the evaluation of the MSE of 0/1 binary fields [see Casati et al. (2004), Eq. (11)] and is intuitive when comparing the forecast and observation binary field difference to their corresponding contingency table images (Figs. 1e and 1f). The random binary MSE is then estimated from the sum of the estimates of b/n and c/n for a contingency table obtained by random chance, with sample climatology s and the frequency bias index $B = r/s$ equal to those of the original binary fields. By applying Murphy and Winkler's (1987) factorization and Bayes's theorem, the joint probabilities estimated by b/n and c/n can be expressed as the product of marginal and conditional probabilities (e.g., see Jolliffe and Stephenson 2003; Wilks 2006). Conditional probabilities for random chance are equal to the unconditional probabilities. Then, b/n and c/n are estimated by the products of the marginal probabilities solely. The expected value of the MSE for a biased random binary forecast and the observation fields is then estimated by

$$\mathrm{MSE}_{u,\mathrm{rand}} \sim B \cdot s \cdot (1 - s) + s \cdot (1 - B \cdot s). \tag{4}$$

The IS skill score is defined with the standard skill score definition (as in Jolliffe and Stephenson 2003; Wilks 2006). It is based on the $\mathrm{MSE}_{u,l}$ and uses random chance as the reference forecast. The binary random MSE estimated by Eq. (4) is equipartitioned across the $L + 1$ scales to obtain the IS skill score:

$$\mathrm{SS}_{u,l} = 1 - \frac{\mathrm{MSE}_{u,l}}{\mathrm{MSE}_{u,\mathrm{rand}}/(L + 1)}. \tag{5}$$

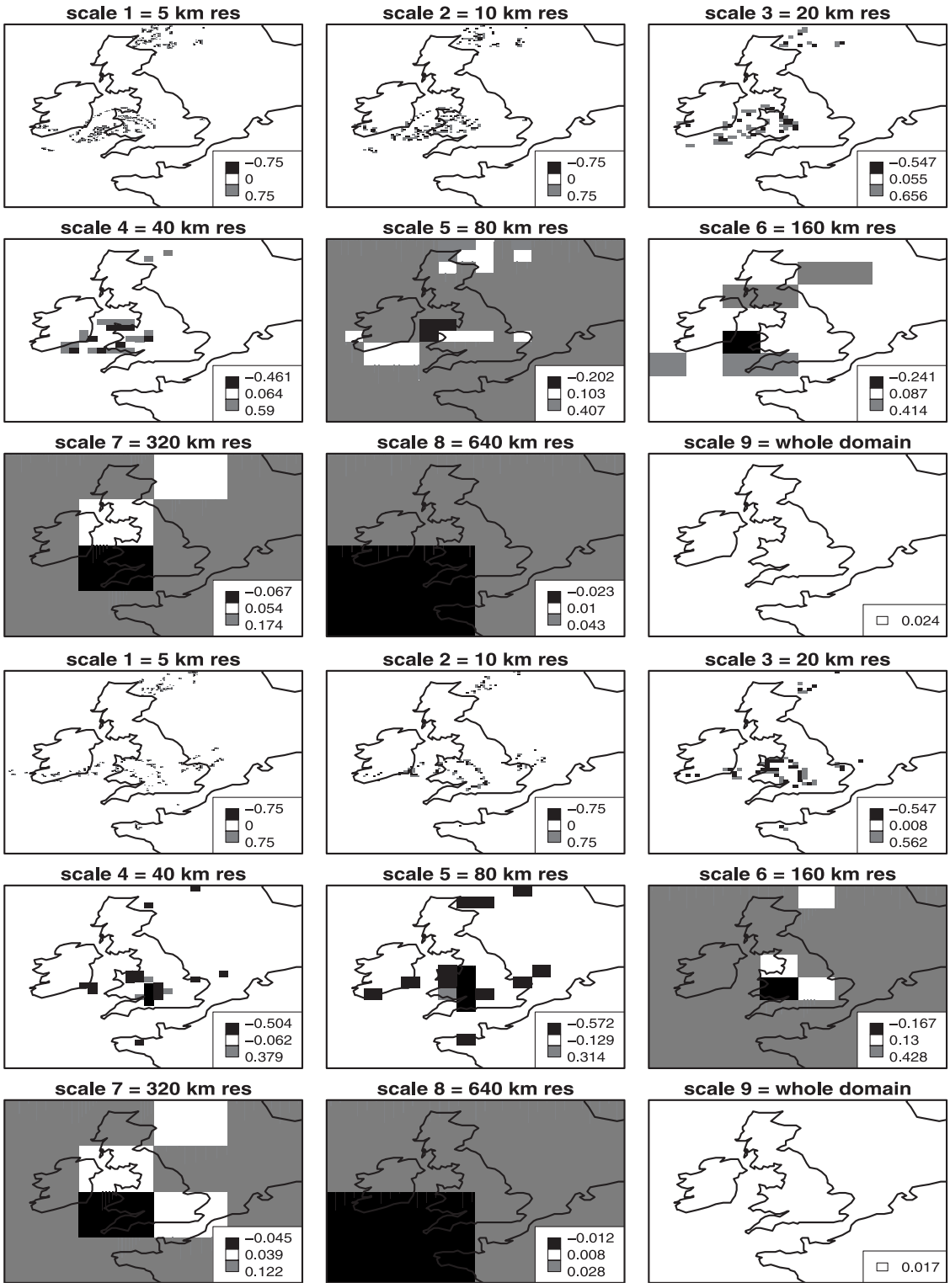

FIG. 2. Wavelet scale components of the (top) binary forecast and (bottom) analysis for the NIMROD case study, for a threshold of 1 mm h^{-1}.

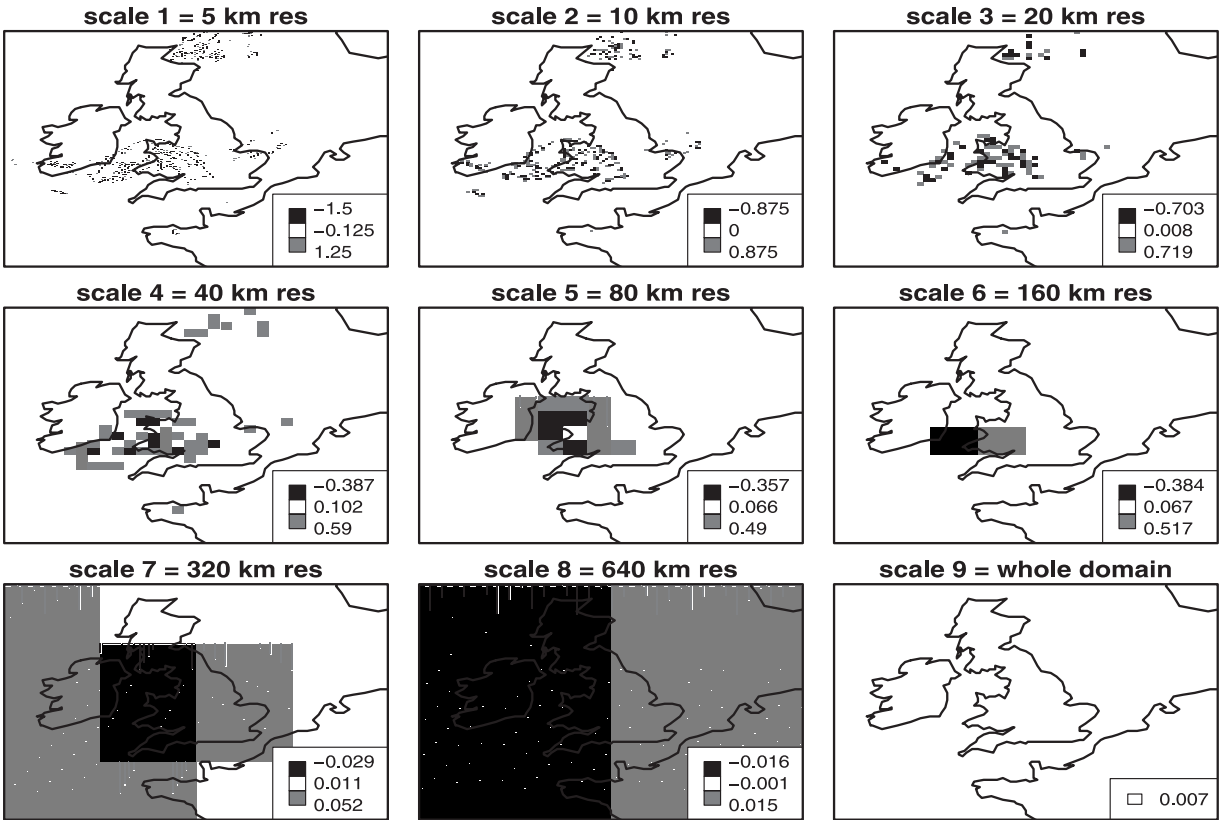

FIG. 3. Wavelet scale components of the binary field difference for the NIMROD case study, for a threshold of 1 mm h^{-1}.

Figure 5 shows the IS skill score for the NIMROD case study. The IS skill score evaluates the forecast skill as a function of the precipitation intensity and the spatial scale of the error. Positive values of the IS skill score are associated with a skillful forecast, whereas negative values are associated with no skill. Usually, large scales exhibit positive skill since large-scale events, such as fronts, are well predicted, whereas small scales exhibit less skill, since small-scale events, such as convective showers, are less predictable. The smallest scales associated with the highest thresholds usually exhibit the worst skill. For the NIMROD case illustrated, the no-skill to skill transition scale is 40 km. However, the 160-km scale exhibits negative skill for the thresholds from ½ to 4 mm h^{-1} due to the 160-km storm, which is displaced almost its entire length.

From Eq. (5) and Eqs. (1), (3), and (4), it is easy to show that for each threshold u the average of the scale components of the IS skill score is equal to the Heidke skill score (see section 3.3 of Casati et al. 2004). The IS skill score can be interpreted as a scale decomposition of the Heidke skill score (HSS). The intensity-scale technique, in this respect, bridges categorical and scale-separation (or scale decomposition) verification approaches.

b. The energy

To assess the bias on different scales and compare forecast and observed scale structures, the energy (En) and its percentages (En%) are evaluated. The average energy per cell (hereafter energy) of a gridded field X is the average of the field gridpoint squared values:

$$\mathrm{En}(X) = \frac{1}{n}\sum_{i=1}^{n} x_i^2. \tag{6}$$

As for the MSE, we denote the energy of the original binary fields as En_u, and the energy of the scale components of the binary fields as $\mathrm{En}_{u,l}$, to indicate their dependency on the threshold u and scale l. The energy of the scale components of the binary forecast and observed fields provides feedback on the number of events present in the forecast and observation, as a function of threshold and scale. Figures 6a and 6b show $\mathrm{En}_{u,l}$ for the NIMROD case study forecast and analysis: small thresholds are associated with high energy, since many events exceed the threshold, whereas large thresholds are associated with low energy, since few events exceed the threshold. Note, especially for the analysis, the higher energy at the 160-km scale and for thresholds between

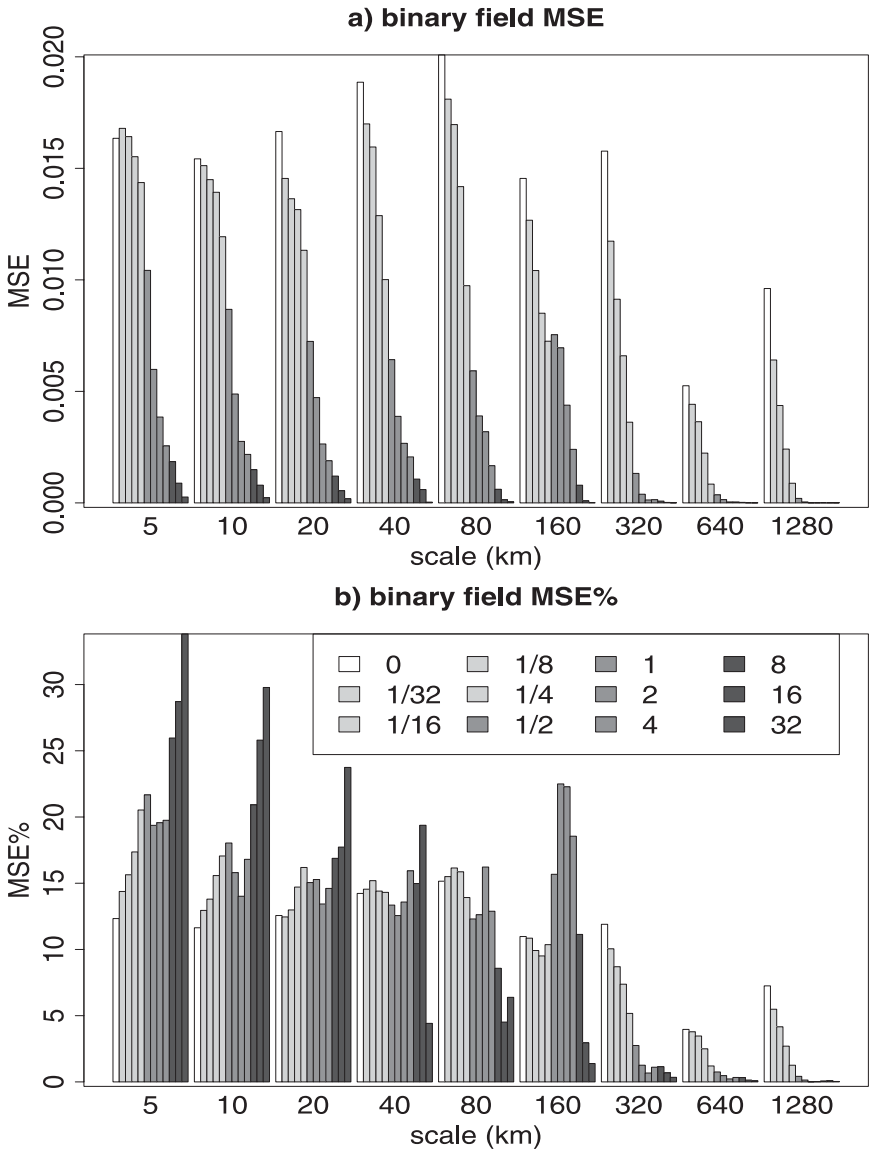

FIG. 4. Binary field (a) MSE and (b) MSE percentage for the NIMROD case study. Bars of different gray shades correspond to increasing precipitation thresholds (mm h^{-1}), as indicated in the legend.

½ and 4 mm h^{-1}, which is associated with the intense 160-km storm.

Comparison of the forecast and observation $En_{u,l}$ provides feedback on the bias on different scales and for each threshold. The energy bias can be assessed by the energy difference, the energy ratio, and by the energy relative difference. The energy relative difference is defined as the difference between forecast (F) and observation (O) energies normalized by their sum:

$$\text{En rel diff} = \frac{[\text{En}(F) - \text{En}(O)]}{[\text{En}(F) + \text{En}(O)]}. \qquad (7)$$

By dividing the numerator and denominator of this equation by the observation energy, it is easy to show that the energy relative difference is equal to the forecast and observation energy ratio centered on zero and scaled, in order to range between −1 and 1. Positive values of the energy relative difference indicate overforecasting and negative values indicate underforecasting. The perfectly unbiased forecast would have zero energy relative difference.

In this study, for quantitative precipitation forecasts, the energy relative difference is preferred over the energy difference or ratio for the following reasons:

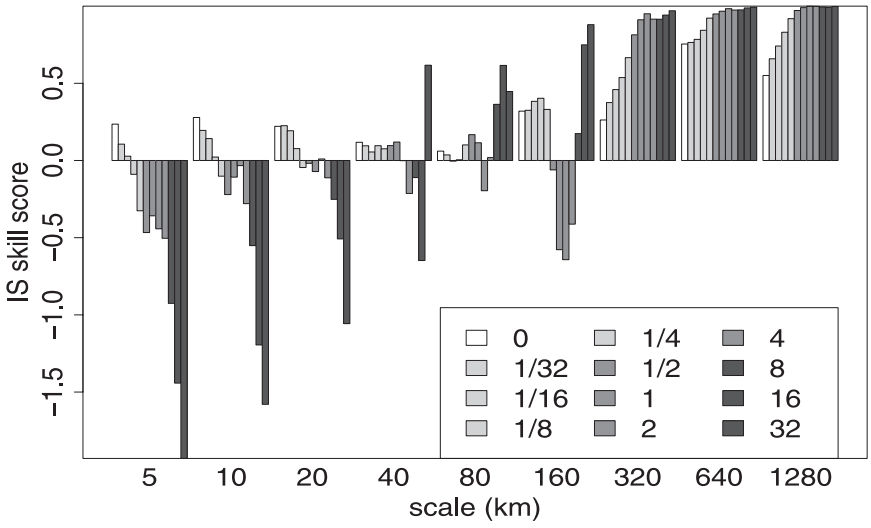

FIG. 5. Intensity-scale skill scores for the NIMROD case study. Bars of different gray shades correspond to increasing precipitation thresholds (mm h^{-1}), as indicated in the legend.

1) The energy difference tends to overpenalize (underpenalize) differences that are large (small) in absolute value simply because the two energy amounts are large (small). Instead, the energy relative difference accounts for the difference between the forecast and observation energies relative to their magnitude.
2) The energy ratio compares the forecast and observation energies, accounting for their relative magnitudes. However, the ratio approaches infinity as the denominator (observation energy) approaches zero. The energy relative difference is equal to the centered and scaled energy ratio; therefore, it retains the same type of information but has a finite range.

The energy relative difference, therefore, enables one to compare biases associated with low- and high-precipitation thresholds (associated with larger and smaller energy amounts) on a common scale.

Figure 6c shows the forecast and analysis $\mathrm{En}_{u,l}$ relative difference for the NIMROD case study. The forecast has too much energy at small thresholds, especially for the large scales, due mainly to the overforecasting of drizzle in the north (off the Scottish east coast) and to the west of the 160-km storm (see Fig. 1). On the other hand, the energy relative difference exhibits underforecasting for high thresholds, consistent with the lack of intense precipitation in the NIMROD forecast. Note that these results agree with the results deduced using the recalibration function in Casati et al. (2004). The scale of 160 km exhibits a particularly pronounced underestimation for the thresholds between 2 and 8 mm h^{-1}, since the 160-km storm in the analysis is characterized by larger values than those forecast.

The energies of forecast and observation binary fields are related to categorical statistics. The energy of the observation binary field, $\mathrm{En}_u(O)$, is equal to the sample climatology $s = (a + c)/n$ for the contingency table obtained with the same threshold. Similarly, the energy of the forecast binary field, $\mathrm{En}_u(F)$, is equal to $r = (a + b)/n$. These relations can be easily shown by evaluating the energy of the 0/1 binary fields via Eq. (6), and noting that $x_i = x_i^2$ for binary 0/1 gridpoint values. The relations are intuitive when comparing binary forecast and analysis fields to their corresponding contingency table image (Figs. 1c,d, and 1f). The En_u relative difference is equal to the statistics obtained by centering and scaling the frequency bias index: $(B - 1)/(B + 1)$.

To focus on the forecast and observation scale structure solely, the threshold dependence of the energy is eliminated by normalization. As for the MSE [see Eq. (1) and related discussion], because of the orthogonality of the discrete wavelet transform, the additive properties of the scale components transfer to the energy. The sum of the energies of the scale components of the binary field is then equal to the energy of the original binary field:

$$\mathrm{En}_u = \sum_{l=1}^{L+1} \mathrm{En}_{u,l} \tag{8}$$

This enables one to calculate, for each threshold, the percentage of the total En_u that each scale contributes:

$$\mathrm{En}\%_{u,l} = \left(\frac{\mathrm{En}_{u,l}}{\mathrm{En}_u}\right) \times 100. \tag{9}$$

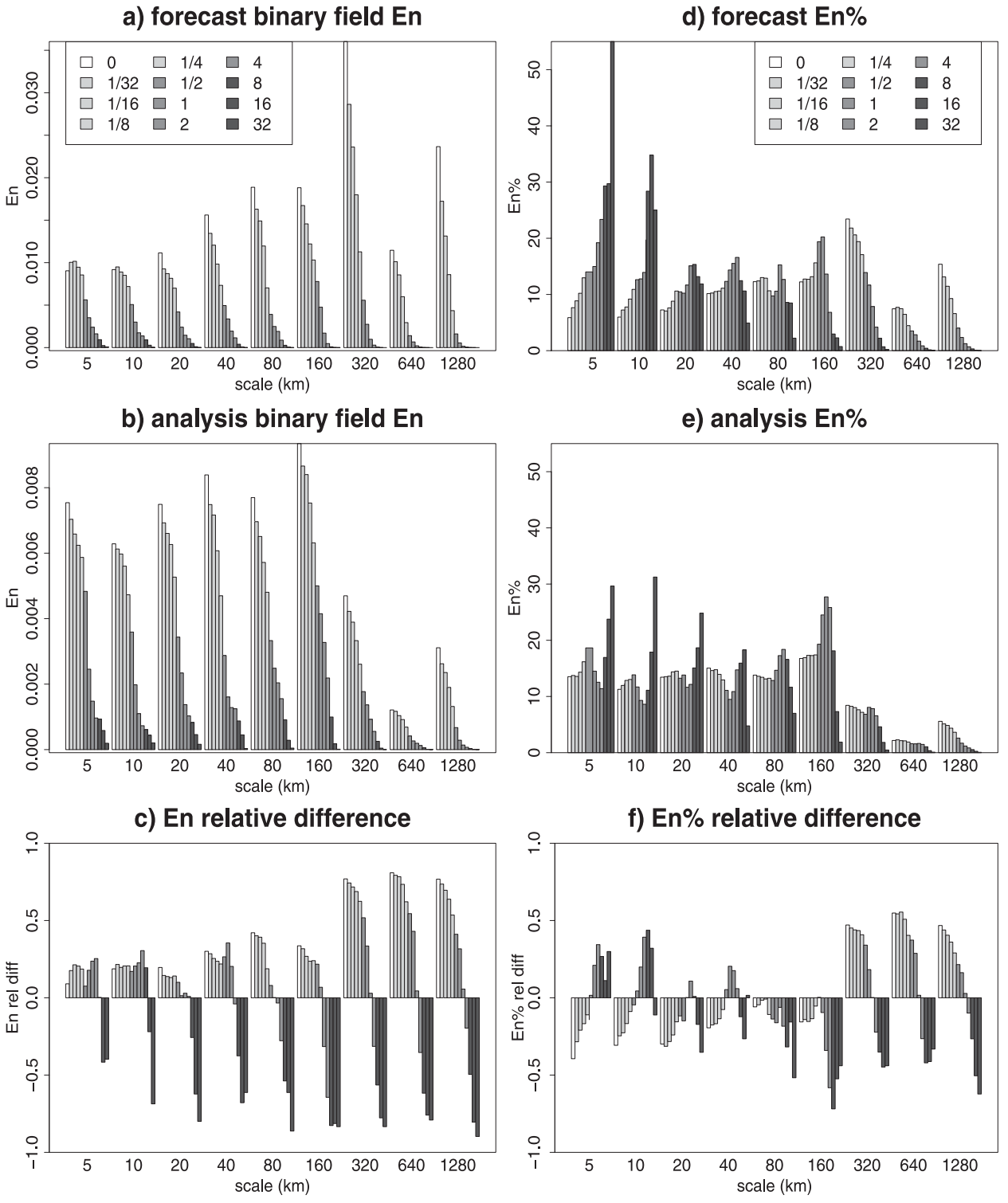

FIG. 6. NIMROD (a) forecast and (b) analysis energy, and (c) their relative difference. NIMROD (d) forecast and (e) analysis energy percentage, and (f) their relative difference. Bars of different gray shades correspond to increasing precipitation thresholds (mm h^{-1}), as indicated in the legend.

The energy percentages provide feedback on how the events are distributed across the scales and, therefore, carry information on the scale structure of the forecast and observation fields. Figures 6d and 6e show the $\mathrm{En}\%_{u,l}$ for the NIMROD case study. On large scales, a larger energy percentage is associated with low thresholds (and less with high thresholds), since large-scale features are usually characterized by low intensities (e.g., stratiform precipitation). On the other hand, on small scales the energy percentage is larger for high thresholds (and smaller for low thresholds), since small-scale features are in general associated with intense events, such as convective cells or showers. Similarly, looking across the scales, for high thresholds most of the energy percentage is concentrated on small scales, since high thresholds identify mainly small-scale intense events. For low thresholds, the observed energy percentage is more or less constant up to the 160–320-km scales (size of the observed storm), and then it drops, since the low thresholds identify all large- and small-scale events up to the largest feature size. For the NIMROD case study illustrated, the analysis exhibits large energy percentages on the 160-km scale and for thresholds between ½ and 4 mm h^{-1}, associated with the intense 160-km storm. The forecast energy percentages associated with the 160-km-scale storm are smaller than for the analysis (especially for the larger thresholds, such as $u = 2$, 4, and 8 mm h^{-1}), due to the lower intensities predicted within the storm. On the other hand, for small thresholds the forecast energy percentages on the 320-km scale are large, due to the two broad regions of overforecast drizzle.

Comparison of the forecast and observation energy percentages enables the comparison of the forecast and observation scale structures. The scale structure is assessed by the relative difference of $\mathrm{En}\%_{u,l}$ (Fig. 6f). For small thresholds, the NIMROD forecast overestimates the number of large-scale events and underestimates the number of small-scale events, due again to the drizzle overforecasting. For large thresholds the forecast underestimates the number of large-scale events. This underestimation is particularly pronounced for the 160-km-scale storm at large thresholds (e.g., $u = 2, 4, 8$, and 16 mm h^{-1}): in fact, for these thresholds the analysis still exhibits a large-scale coherent feature (see Fig. 1), whereas the forecast, because of the underestimation of the storm's precipitation intensities, no longer exhibits a large spatially coherent feature. On the other hand, the forecast exhibits an intense and narrow precipitation band along the northern edge of the storm, which is not present in the analysis (see Fig. 1): this causes the overforecasting of the energy percentage for small scales (5–10 km) and high thresholds.

Note that the $\mathrm{En}_{u,l}$ and $\mathrm{En}\%_{u,l}$ intensity-scale structures (Fig. 6) are similar to those of the $\mathrm{MSE}_{u,l}$ and $\mathrm{MSE}\%_{u,l}$ (Fig. 4): this is due partially to the dependence of the MSE on the number of events (the more events, the more likely the error). The IS skill score, on the other hand, is less affected by the number of events, and it is not as dependent on the threshold as is the MSE or energy. In fact, the IS skill score accounts for the error relative to the number of events, via the normalization of the MSE with the $\mathrm{MSE}_{u,\mathrm{rand}}$.

The underlying motivation for choosing the energy to assess the bias on different scales, rather than a different measure, arises from the fact that the energy is a mean square statistic defined in a similar fashion to that for the MSE. Then, as for the MSE, the energy is additive across the scales [Eq. (8)], which allows the evaluation of the threshold-independent energy percentages for the assessment of the scale structure. As for the MSE, the energy relates to categorical statistics when applied to binary 0/1 fields. Finally, since the energy is defined coherently with the MSE, this can naturally lead to the definition of alternative IS skill scores based on $\mathrm{MSE}_{u,l}$, $\mathrm{En}_{u,l}(F)$, and $\mathrm{En}_{u,l}(O)$. As an example, an IS skill score similar to the fraction skill score (Roberts and Lean 2008) can be defined for each threshold u and scale l: this would enable the assessment of the skill relative to the amount of forecast and observed events characterizing each separate scale, and not with respect to an equipartitioned random error (which might not reflect the lower predictability of small scales). Note, however, that with this alternative definition the link between HSS and the IS skill score would be lost.

c. *Statistics aggregation on multiple cases*

The IS statistics obtained from multiple model runs can be aggregated. For consistency, the aggregation is performed for cases defined on the same spatial domain, and aggregated statistics are evaluated for the same intensity thresholds. In this study we will show aggregated statistics for the spring 2005 forecast–observation dataset (section 3c). Here, we briefly discuss how to evaluate the IS aggregated statistics.

For each threshold and scale, the aggregated $\mathrm{MSE}_{u,l}$ and forecast and observation $\mathrm{En}_{u,l}$ are obtained simply by averaging the $\mathrm{MSE}_{u,l}$ and $\mathrm{En}_{u,l}$ of all the model runs, for the corresponding threshold u and scale l. The aggregated MSE_u and En_u are obtained from the aggregated $\mathrm{MSE}_{u,l}$ and $\mathrm{En}_{u,l}$ as the sum of their scale components [Eqs. (1) and (8)]. Aggregated $\mathrm{En}\%_{u,l}$ are obtained from the aggregated $\mathrm{En}_{u,l}$ and En_u via Eq. (9). The energy and energy percentage relative differences are defined from the forecast and observation aggregated $\mathrm{En}_{u,l}$ and $\mathrm{En}\%_{u,l}$ via Eq. (7).

The aggregated IS skill score is evaluated by substituting the aggregated $MSE_{u,l}$ in Eq. (5), and by evaluating the $MSE_{u,rand}$ [Eq. (4)] with the aggregated sample climatology and the aggregated frequency bias index. For each threshold u, the aggregated sample climatology $s = (a + c)/n$ is equal to the aggregated observation $En_u(O)$. Similarly, for each threshold u, the aggregated $r = (a + b)/n$ is equal to the aggregated forecast $En_u(F)$. The aggregated frequency bias index, $B = (a + b)/(a + c)$, is then obtained as the ratio of the aggregated r and s, for the corresponding threshold u.

For all the aggregated statistics, 95% confidence intervals (CIs) are obtained by using a bootstrapping technique (Efron and Tibshirani 1993). The aggregated cases are resampled by random selection, with replacement, for 1001 times. Aggregated statistics are evaluated for each resample to obtain a distribution of 1001 elements for each of the IS statistics (for each threshold and scale). The 0.025 and 0.975 quantiles of these distributions are then used to provide an estimate of the 95% CI for the IS statistics, for each threshold and scale. Note that such confidence intervals reflect the variability of the verification statistics within the aggregated case studies.

d. Dyadic domain constraints

The IS technique as proposed in Casati et al. (2004) was applied to spatial forecasts defined over square domains of $2^n \times 2^n$ grid points (*dyadic domain*). The demand of a dyadic domain is related to the dyadic nature of the discrete (orthogonal) wavelet transforms. This constraint, which is unlikely met for most operational forecasts, can be bypassed in different ways. In this study we propose four different approaches to address this issue: padding, cropping, interpolating, and tiling. The choice of which approach to use to tackle the dyadic domain constraint depends on the shape and dimension of the forecast domain, and on the aims of the verification.

1) *Padding*—If the original forecast domain size is just slightly smaller than a dyadic domain, then a dyadic verification domain can be obtained by appending constant values (e.g., zeros, for precipitation) to the forecast domain boundaries. Note that for variables that are highly discontinuous in space, such as precipitation, padding with zero would not dramatically affect the field distribution (which is characterized by a large number of zero values anyway) and/or the field scale properties. On the other hand, for smooth variables (such as temperature) padding with a constant value would be more problematic, since the distribution and spatial properties of the field would be changed. Padding therefore is advisable only for the former type of variables. Note also that padding might artificially increase the skill, since correct nonevents are added to the forecast and observation fields. Padding is advisable therefore only for small areas.
2) *Cropping (masking)*—The IS verification can be performed on a dyadic domain embedded within the original forecast domain. This approach is suggested if the domain size is just slightly larger than a dyadic domain (boundaries are often disregarded in verification procedures anyway since they are frequently affected by instabilities and other boundary effects), or if it is desirable to reduce the verification domain to a subregion of interest within the forecast domain [e.g., where the observational spatial coverage is more reliable, such as in Casati et al. (2004)].
3) *Interpolating (regridding)*—If the domain dimensions are of similar order, the forecast and observations can be regridded into a dyadic domain by interpolation. Different types of interpolation are advised for different variables, which are characterized by different distributions and field characteristics; for example, nearest-neighbor interpolation or linear interpolation to a denser grid are often advised for precipitation, because of its discontinuous nature and to preserve peak values (mass-conservative interpolations are also often used); cubic-spline interpolation is advised for smoother variables, such as temperature. Note that the interpolation changes the physical scale dimensions and alters the original field values.
4) *Tiling*—This approach does not involve any domain reduction, expansion, or altering of values by interpolation, and it is the most robust of the approaches proposed when applied to a single forecast. Tiling performs the IS verification on dyadic tiles with dimensions equal to those of the largest $2^n \times 2^n$ gridpoint tile that fits in the forecast domain. The tiles are shifted within the forecast domain in order to cover it, allowing tiles to overlap. The IS verification statistics obtained for each tile are then aggregated as described in the previous section. Note that with this approach the center of the domain will be sampled more than the boundary grid values (which can be desirable sometimes). The tiling approach smoothes out the effects due to the discreteness of the wavelet transform support. Tiling also enables one to provide confidence intervals for the IS verification statistics for a single case study. Note, however, that the confidence intervals obtained via the tiling reflect the uncertainty of the statistics due to the discrete support of the wavelet transforms (and not their variability within different aggregated case studies). Finally, when aggregating multiple case

studies or model runs, the effects due to the discreteness of the wavelet transform are naturally eliminated as a result of the movement of weather features. In fact, the precipitation features assume different positions within the discrete wavelet support for each aggregated case, which results in the same effect as moving the tiles for a single case. Therefore, fewer tiles are needed for aggregated cases; for large aggregations, one tile position is probably sufficient.

All the case studies used in the Intercomparison of Spatial Forecast Verification Methods are defined over spatial domains of 501 × 601 grid points. For the geometric cases (section 3a), the IS verification results are obtained by tiling. Statistics evaluated on 201 tiles of 512 × 512 grid points are aggregated. The tiles are randomly positioned, so that the forecast and reference geometric shapes assume different positions with respect to the discrete wavelet dyadic support, but are still entirely contained in the domain and with the same relative positions with respect to each other. Tiles with values beyond the original 501 × 601 gridpoint domain are padded with zeros. Some of the results obtained from single tiles are also shown, in order to illustrate the sensitivity of the verification statistics to the discrete wavelet support.

For the synthetically perturbed case study (section 3b), the original 501 × 601 gridpoint domain is first padded with a narrow stripe of zeros (11 × 601 grid points), in order to obtain a domain of 512 × 601 grid points. The IS verification statistics are then obtained by tiling with five tiles of 512 × 512 grid points, with origins at the columns 1, 22, 45, 67, and 90. These five tiles, with positions varying in the x direction solely, are chosen in order to cover entirely the precipitation domain while minimizing the zero-padded areas (tiles with positions that also vary in the y direction would require more padding, which could add artificial skill). Moreover, fewer tiles are used for the perturbed case than for the geometric cases, since for more realistic precipitation fields the IS statistics exhibit less sensitivity to the wavelet support position (more discussion follows in sections 3a and 3b). As for the geometric cases, results obtained by tiling are compared to those obtained from single tiles in order to illustrate how the tiling procedure reduces the effects due to the discrete wavelet support.

For the spring 2005 dataset (section 3c), the domain is first padded with zeros in the same fashion as for the synthetically perturbed case. The three approaches—cropping, interpolating, and tiling—are then applied and compared, through the analysis of the distinct behaviors of the IS verification statistics for the aggregated case studies. The cropping is performed differently for each of the nine cases, depending on the position of the major precipitation features: the region (either to the west or to the east of the domain) of 512 × 89 grid points with less precipitation is masked out prior to performing the IS verification. Note that such a case-by-case selected cropping procedure is performed here solely to better illustrate the cropping effects on the IS statistics (see section 3c), and it is not recommended for operational practice. A nearest-neighbor interpolation is performed to regrid the spring 2005 case studies on a dyadic 512 × 512 gridpoint domain; note that this interpolation eliminates columns of values (evenly spaced across the east–west direction) from the original fields. Tiling is performed as for the perturbed case study, with five tiles of 512 × 512 grid points with origins at columns 1, 22, 45, 67, and 90. Results obtained with a larger number of tiles were similar to those found with these five tiles (not shown).

3. Results

a. The geometric cases

The IS skill score and energy are evaluated for the geometric case studies (Fig. 1 in Ahijevych et al. 2009), for thresholds of 12.7 and 25.4 mm. The lower threshold of 12.7 mm identifies the entire elliptic feature (low and high intensities), whereas the higher threshold of 25.4 mm isolates the high-intensity core embedded in the low-intensity feature. Figure 7 shows the IS skill score obtained by tiling for the geometric case studies. For all the cases, the skill is negative at the dominant scales of the features and their displacements. For the higher threshold, the negative skill is associated with smaller scales than for the lower threshold, since the higher threshold isolates smaller-scale features.

Figures 7a and 7b provide feedback on the sensitivity of the skill score to displacements of the elliptic feature. As the distance grows, the negative skill shifts to larger scales, for both the low and high thresholds: the IS skill score is sensitive to the displacement error. Figures 7c and 7e provide feedback on the sensitivity of the skill score to errors in the extent of the feature, for the same displacement. As the feature extent gets larger, the negative skill score shifts to larger scales, for both the low and high thresholds: the IS skill score is sensitive to feature size errors. Note also that despite the larger bias, the skill score in Fig. 7e is less negative than that in Fig. 7c for the low threshold, because the forecast and reference ellipses overlap. The IS skill score does not separate the different sources of forecast errors, such as displacement or feature size, but it is sensitive to them.

Figure 7d shows the IS skill score for the elliptic feature, which is displaced and deformed to have a horizontal

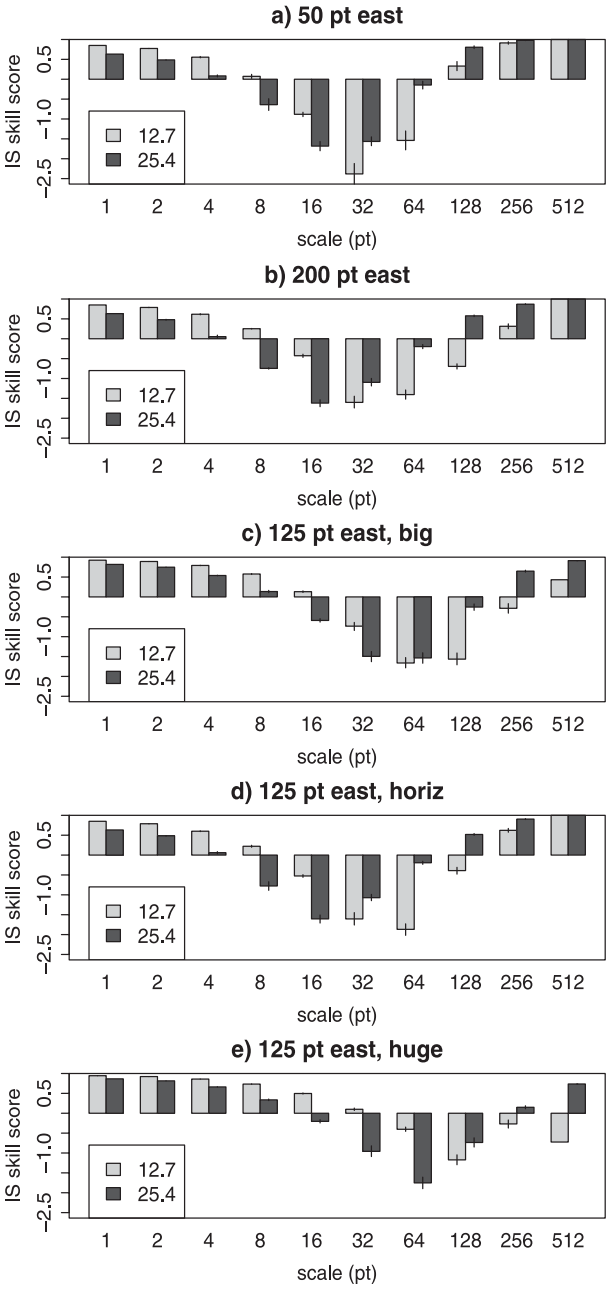

FIG. 7. Intensity-scale skill score obtained by aggregating multiple tiles, for the geometric cases. Bars of different gray shades correspond to different thresholds (mm). Segments at the bar extremities indicate the bootstrap 95% confidence intervals for the aggregated tiles.

major axis. This case aims to test the sensitivity of the verification technique to the feature orientation. The error scale structure (i.e., the scale components of the binary field difference) and the $MSE_{u,l}$ for this case are in between those of the ellipsoids displaced 50 and 200 grid points. In fact, the forecast and reference ellipsoids touch (as for the ellipsoid displaced 50 grid points) but, because of the horizontal orientation, the wavelet components of the binary field difference extend to larger scales (as for the ellipsoid that is displaced 200 grid points). The IS skill score for this case is similar to that of the ellipsoid displaced 200 grid points. The sensitivity of the IS skill score to the feature orientation is not as well defined as the sensitivity to the displacement and feature size.

Note that for the geometric cases the IS skill score is positive not only for very large scales (which are sufficiently large to remain unaffected by the ellipses' displacements), but also for the smallest scales: the smallest scales are skillful because there are no small-scale events (the ellipses are smoother than the grid resolution), and therefore little error is associated with these scales. On the other hand, for these same geometric cases, neighborhood verification techniques [e.g., the fraction skill score; see Mittermaier and Roberts (2010); Ebert (2009)] exhibit negative skill on small scales, and the skill becomes positive only when the neighborhood size is sufficiently large to encompass both the reference and the displaced ellipses. These very different results are due to the different definitions of ''scale'' for the scale-separation and neighborhood verification approaches. For scale-separation methods, such as the IS technique, the scales are obtained with a single-band filter: these approaches are therefore able to isolate scale-dependent errors and assess the skill separately, for each individual scale. Neighborhood methods, on the other hand, are based on a low-bandpass filter (i.e., smoothing): as the neighborhood size (or scale) increases, forecast and observation fields are filtered and the exact space–time matching requirements become more and more relaxed. For neighborhood approaches, then, the skill increases with increasing scales because of their intrinsic definition: these methods do not separate the skill by scale, but assess the resolution for which, by smoothing, the wanted skill is achieved.

Figure 8 shows the energy obtained by aggregating multiple tiles for three of the geometric cases characterized by forecast elliptic features of different size. From the different ranges of the y axes, one notes that the energy for the lower threshold (Fig. 8, left) is larger than the energy for the higher threshold (Fig. 8, right), since the energy is proportional to the number of grid points exceeding the threshold.

Figures 8a and 8d show the energy of the geometric case in which the forecast ellipse is rotated and displaced, but its size is unchanged with respect to the reference ellipse. The forecast and observation energies are identical; in fact, the energy describes the spectral structure of the forecast and observation fields independently of the

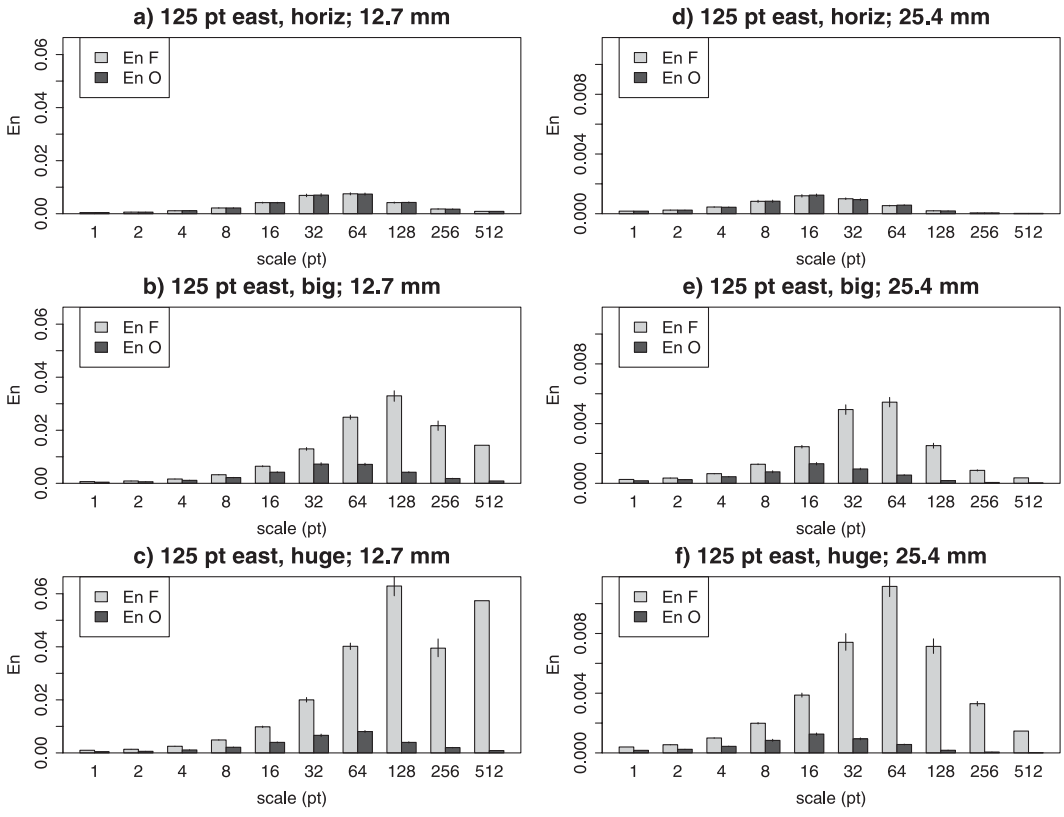

FIG. 8. Forecast (light gray bars) and observed (dark gray bars) energy for thresholds of (left) 12.7 and (right) 25.4 mm obtained by aggregating multiple tiles, for geometric cases with forecast elliptic features of different sizes. Segments at the bar extremities indicate the bootstrap 95% confidence intervals for the aggregated tiles.

features' positions and/or their vertical versus horizontal orientations. For both the forecast and observations, the largest energy identifies a medium- to large-scale feature for the low threshold (32–64 grid points, corresponding to the larger ellipse), and a small- to medium-scale feature for the high threshold (16–32 grid points, corresponding to the small-scale core embedded in the larger ellipse). Because of the energy independence from the feature positions and their x–y orientation, all energy-based statistics for the two geometric cases with forecast elliptic features of the same size as the reference ellipse, but displaced (50 and 200 grid points to the east), exhibit the same behavior as this illustrated case and are, therefore, not shown.

Figures 8b,e and 8c,f show the energies for the geometric cases in which the ellipse sizes are set larger and larger with respect to the reference (hereafter these are referred to as the *big* and *huge* ellipses, respectively). The larger feature sizes are evident in the distribution of the energy: for the low threshold (Figs. 8b and 8c), the largest amount of energy identifies a 128-gridpoint scale feature for the big ellipse, and a 128–256–512-gridpoint scale feature for the huge ellipse; for the high threshold (Figs. 8e and 8f), the intense cell embedded in the less intense ellipse corresponds to large energy amounts at the scale of 32–64 grid points for the big ellipse, and at the scale of 64 grid points for the huge ellipse. Note that for both the low and high thresholds, the largest amount of energy (in absolute value) is associated with the largest (huge) ellipse (Figs. 8c and 8f).

Figure 9 shows the forecast and observation energy differences obtained by tiling for the geometric cases shown in Fig. 8. As expected, the geometric case with a forecast ellipse of the same size as the reference ellipse exhibits no bias (Fig. 9a). On the other hand, the geometric cases with larger forecast ellipses exhibit positive bias in correspondence to the scales of the overforecast features (128 and 32–64 grid points in Fig. 9b; 128–256–512 and 64 grid points in Fig. 9c), for the corresponding threshold. Note that for the geometric cases the bias is assessed by the energy difference, instead of the relative difference. This is done to illustrate an example in which the bias is proportional to the energy magnitude; a comparison of Figs. 9b and 9c shows that the largest bias is associated with the largest feature. The energy difference is appropriate for verification applications that

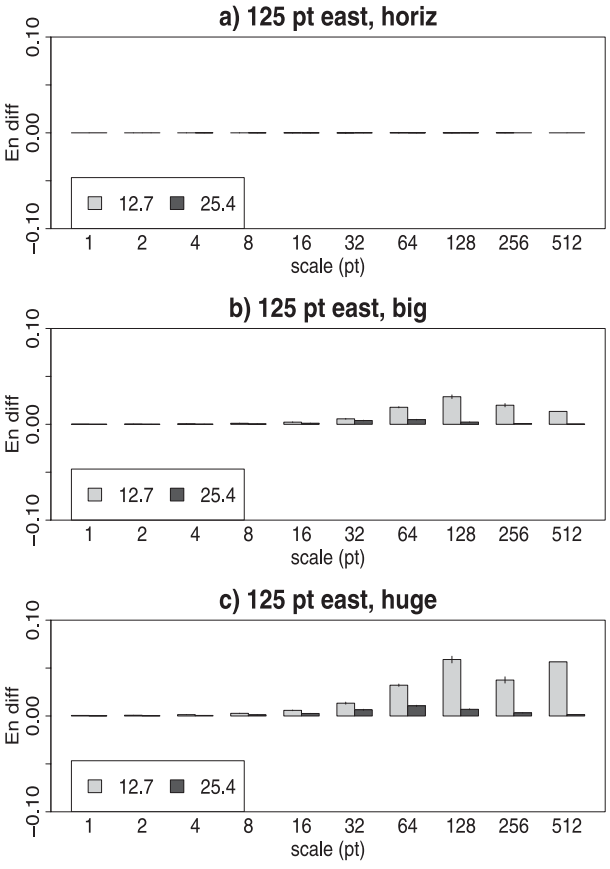

FIG. 9. Energy difference obtained by aggregating multiple tiles, for geometric cases with forecast elliptic features of different sizes. Bars of different gray shades correspond to different thresholds (mm). Segments at the bar extremities indicate the bootstrap 95% confidence intervals for the aggregated tiles.

do not want to excessively penalize small biases associated with small energies, but want to focus on events with large energies (and large biases).

Figure 10 shows energy percentages (En%) obtained by tiling for the geometric cases shown in Fig. 8. The energy percentages are normalized and, therefore, provide information on how the energy is distributed across the scales independently of the energy magnitude; the bars of all the panels of Fig. 10 (as opposed to Fig. 8) therefore have similar ranges, despite being associated with different thresholds and/or features of different sizes. For each case study, and for each of the two thresholds, the En% identifies the same features that are identified by the energy. The differences between the forecast and reference feature sizes (i.e., the differences in the scale structure of the fields) are then detected by the shifts of the energy percentages: for the low threshold, the En% shifts from medium to large and very large scales (Figs. 10b and 10c); for the high threshold, the En% shifts from small to medium scales (Figs. 10e and 10f). For the case with the forecast elliptic feature that is identical to the reference ellipse, the energy percentages exhibit no shift (Figs. 10a and 10d). The scale shifts are more evident when comparing the forecast and observation energy percentages, rather than their energies (cf. Figs. 8 and 10).

The En% difference shown in Fig. 11 enables us to assess the scale structure independently of the bias associated with the thresholds (the bias has, in fact, been removed by the normalization of the energy percentages). The case with the forecast elliptic feature that is identical to the reference ellipse has ''perfect'' scale structure (Fig. 11a), whereas the cases with larger features (Figs. 11b and 11c) overforecast the large scales. The medium scales (associated with the reference ellipse) are correspondingly underforecast. As expected, the under- and overforecasting for the high threshold occurs at smaller scales than for the low threshold (since the high threshold corresponds to the smaller ellipses), and the overforecast for the huge ellipse (Fig. 11c) stretches to larger scales than for the big ellipse (Fig. 11b). Note that, for a given threshold, the En% differences on the different scales add up to zero. In fact, each overforecast for a particular scale is compensated by the underforecasting at other scales.

Figure 12 shows the energy obtained for a single tile, for the three geometric cases with forecast elliptic features identical to the reference ellipse. For both high and low thresholds, the energies differ only because of the positions of the ellipses with respect to the discrete wavelet transform support. Tiling removes this effect, so that both the reference ellipse and the three forecast ellipses exhibit the same energy (e.g., Figs. 8a and 8d). In general, all the IS verification statistics are sensitive to the discreteness of the wavelet support, and tiling and aggregation help to reduce the wavelet discrete support effects. Note that, for the geometric cases illustrated, these effects are not completely eliminated. As an example, the energy percentages for the geometric case with the forecast elliptic feature identical to the reference ellipse are not identical (Figs. 10a and 10d), and their difference departs slightly from zero (Fig. 11a). However, these differences are not significant and their 95% confidence intervals always include zero. To entirely eliminate the effects due to the discrete wavelet support, tiling should be performed for each grid point of the dyadic domain (512×512 tiles). However, this process is computationally very expensive; a tiling coverage that eliminates with statistical significance the discrete wavelet support effects is usually reached with randomly positioned tiles equal in number to the size of the dyadic domain (e.g., 512 tiles for a 512×512 grid

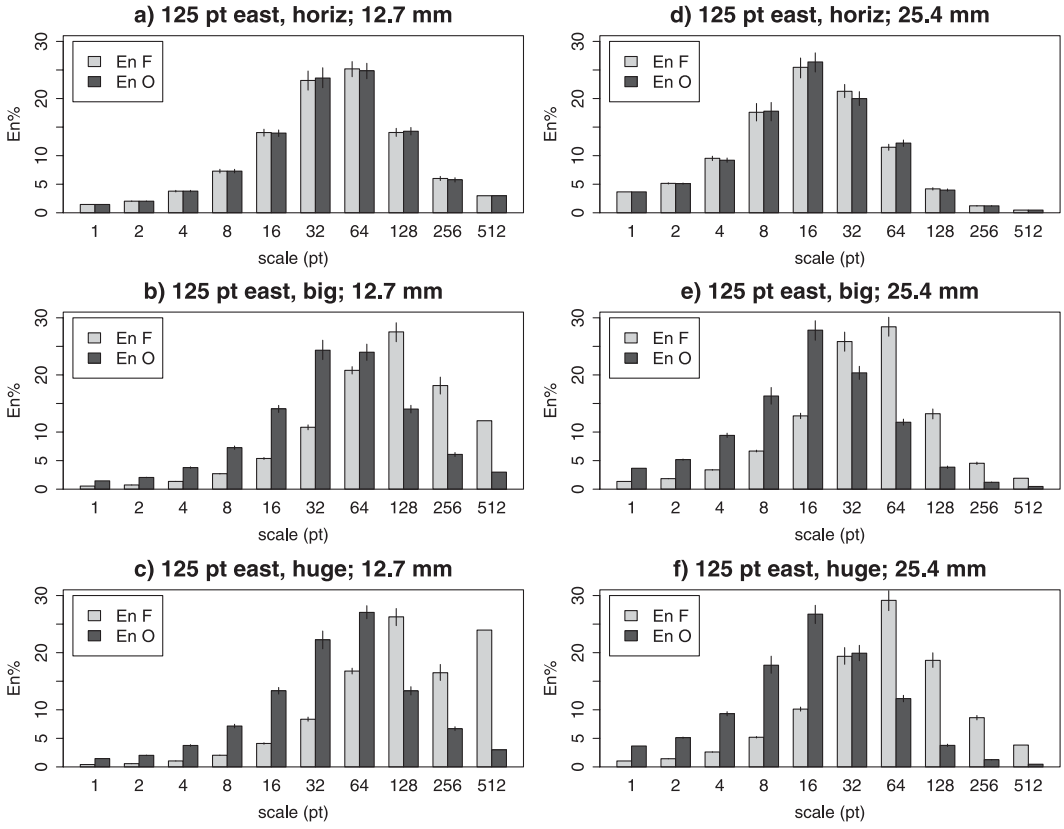

FIG. 10. Forecast (light gray bars) and observed (dark gray bars) energy percentage for thresholds of (left) 12.7 and (right) 25.4 mm obtained by tiling, for geometric cases with forecast elliptic features of different sizes. Segments at the bar extremities indicate the bootstrap 95% confidence intervals for the aggregated tiles.

point domain), since in this way all the positions of the dyadic support in the x and y directions are expected to be covered. For the geometric cases illustrated, fewer tiles are used (201 tiles) to meet the algorithm constraints of preserving the ellipses' relative positions while not cutting through them. In fact, each tile must entirely contain the rectangle of size $\Delta x \times \Delta y$ that embeds both reference and forecast ellipses. Therefore, tiling can be performed with a maximum of $(512 - \Delta x) \times (512 - \Delta y)$ tiles. The square root of this product provides, then, an estimate of the number of tiles randomly positioned in order to cover all of the x and y positions of the dyadic support. Note that 201 tiles are sufficient for obtaining (with 95% confidence) identical energies and zero energy difference for the geometric cases with ellipses of the same size. For more realistic precipitation fields, fewer tiles are needed to eliminate the discrete wavelet support effects (see sections 3b and 3c). This is possibly due to the characteristics of the Haar wavelet filter, which efficiently represents highly discontinuous (noisy) on and off fields, such as precipitation fields (Casati et al. 2004).

b. The synthetically perturbed case

Figure 13 shows the IS skill score obtained by tiling for the case study synthetically perturbed in order to have different displacement errors. The synthetically perturbed case is shown in Fig. 3 of Ahijevych et al. (2009), and the displacements errors correspond to the first five cases listed in Table 4 and described in section 3 of Ahijevych et al. (2009). As the displacement gets larger, the no-skill to skill transition scale (i.e., the scale at which the IS skill score crosses the zero line, from negative to positive) shifts toward larger scales. The IS skill score again exhibits sensitivity to the displacement error. Note also that the shift of the skill-transition scale is important for small thresholds (e.g., Figs. 13a and 13b), whereas high thresholds are less affected (e.g., Fig. 13d). This is partially due to the spatial and physical coherence of the precipitation features and the intrinsic relationship that exists between the feature intensity and scale, so that low thresholds are often associated with large-scale features, whereas high thresholds are associated with small-scale cells. In fact, for the largest threshold the

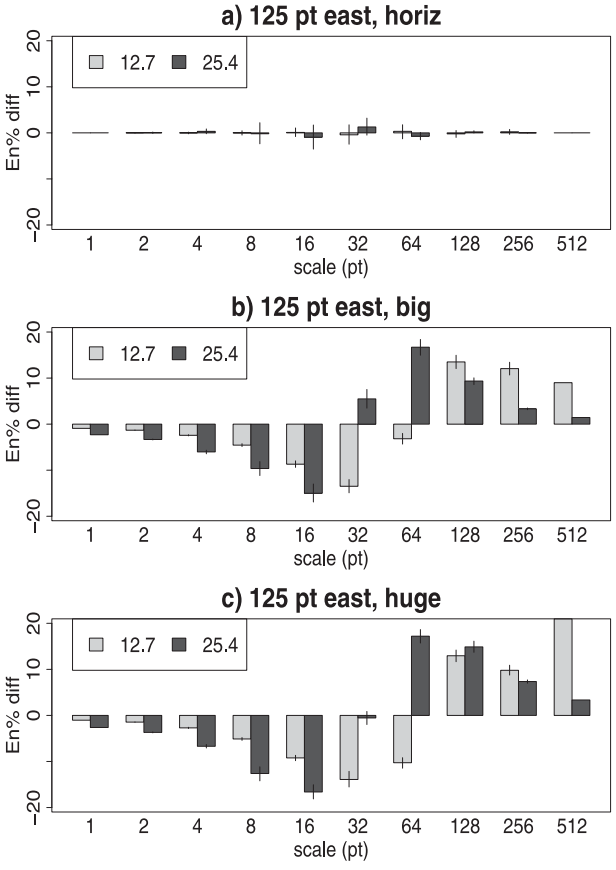

FIG. 11. Energy percentage differences obtained by tiling, for geometric cases with forecast elliptic features of different sizes. Bars of different gray shades correspond to different thresholds (mm). Segments at the bar extremities indicate the bootstrap 95% confidence intervals for the aggregated tiles.

skill remains almost identical, since the largest threshold is associated with very small-scale events (of the size of a grid point or two), which are already completely mismatched in position, from the first small displacement. On the other hand, as the displacement gets larger, the reference and displaced features become more and more separated, and therefore it is not only the intense small-scale cells but also the low-intensity large-scale features that no longer overlap; the skill is then affected more and more for smaller and smaller thresholds. Note that in general for high thresholds the positive skill on large scales is due to the correct zeros.

Figures 14b and 14c show the IS skill scores obtained by tiling for the case study synthetically perturbed to have different intensities, in addition to a displacement error (last two cases listed in Table 4 of Ahijevych et al. 2009). These are compared to the IS skill scores for the perturbed case with the same displacement but no bias (the third case in Table 4 of Ahijevych et al. 2009), which is shown in Fig. 14a. When precipitation intensities are multiplied by a factor of 1.5 (cf. Fig. 14a and 14b), only the highest thresholds (and small scales) are visibly affected. The extent of the precipitation features exceeding these thresholds is slightly larger; therefore, the negative skill associated with the smallest scale shifts to larger scales, and the no-skill to skill transition scale also shifts to a larger scale. On the other hand, when all the precipitation intensities are reduced by the same small quantity (cf. Figs. 14a and 14c), the small thresholds (and large scales) are the most affected. The extent of the large-scale features is reduced, and with it the overlap of the forecast and reference is reduced; therefore, the IS skill score (for small thresholds and large scales) gets worse.

The sensitivity of the IS skill score to the intensity bias error is marginal; the differences between these case studies are better captured by the energy bias. Figure 15 shows the energy obtained by tiling for the reference feature (Fig. 15a), the precipitation feature displaced (Fig. 15b), and the precipitation field synthetically perturbed to have different intensities (Figs. 15c and 15d), in addition to the displacement. Figures 16b–d show the energy relative difference associated with these cases, obtained again by tiling. By multiplying the precipitation intensities by a factor of 1.5, the energy for high thresholds is mostly augmented. This can barely be seen when comparing Figs. 15a and 15c, since the energy for the large thresholds is very small. On the other hand, this overforecasting for high thresholds is well captured by the energy relative difference shown in Fig. 16c. Subtraction of the same small quantity from all the precipitation values reduces the energy, for all thresholds and scales. This is already noticeable by comparing Figs. 15a and 15d, and it is clearly shown by the energy relative difference in Fig. 16d.

The energy relative differences shown in Fig. 16b, comparing the energy of the reference feature (Fig. 15a) and the displaced feature (Fig. 15b), are due solely to the different positions of the precipitation features with respect to the discrete wavelet transform support (i.e., the sensitivity of the statistics to the discrete wavelet support), and should be zero. Only five tiles, with origins at columns 1, 22, 45, 67, and 90, have been aggregated for this case study: the effects of the discrete wavelet support on the statistics are already dramatically reduced, and the energy relative difference is much less noisy and closer to zero than that for single tiles (Fig. 16a). Using 40 randomly positioned tiles removes, at the 95% confidence level, the effects due to the position of the discrete wavelet support (not shown). The same five tiles used for the perturbed case were then chosen for tiling the spring 2005 case studies (section 3c). In fact, the

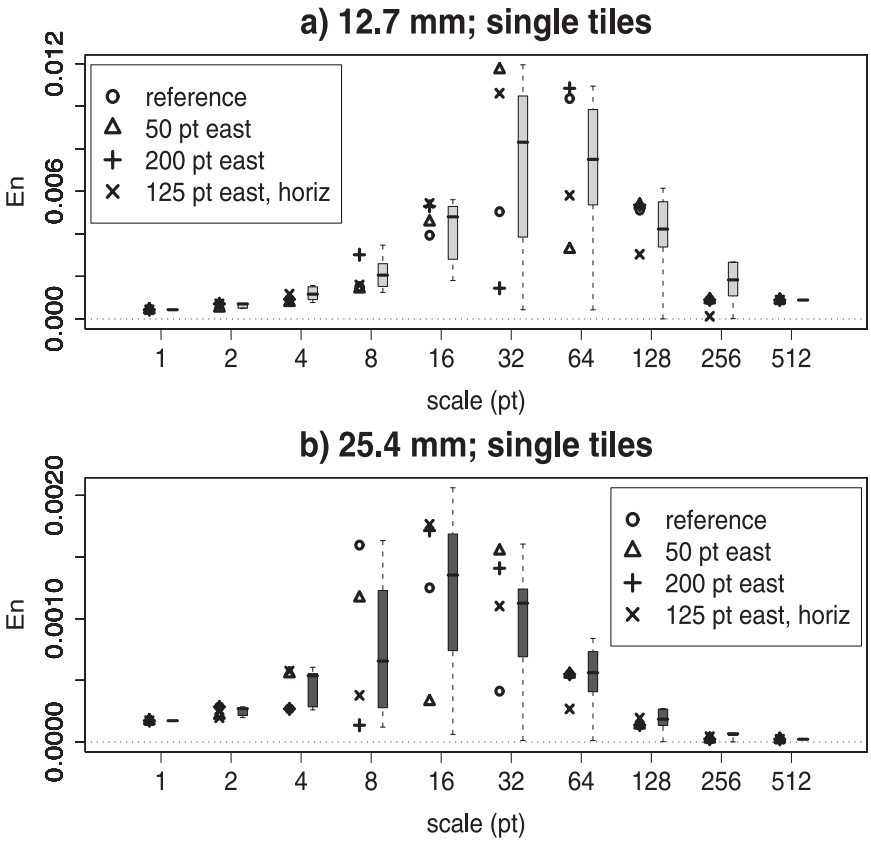

FIG. 12. Forecast and observed energy for thresholds of (top) 12.7 and (bottom) 25.4 mm obtained for a single tile, for the geometric cases with identical elliptic features (symbols). The box plots indicate the distributions of the forecast and observed energies for all of the 201 individual tiles, for the three case studies illustrated.

perturbed case is obtained from one of the spring 2005 cases; therefore, these have similar spatial characteristics. Moreover, the statistics for the spring 2005 datasets are obtained by aggregating nine case studies, for which the positions of the precipitation features with respect to the wavelet support vary case by case. Therefore, five tiles are expected to be sufficient (9 cases $\times$ 5 tiles = 45 $>$ 40 randomly positioned tiles), for the spring 2005 case studies, to eliminate the effects due to the discreteness of the wavelet dyadic support.

c. *The spring 2005 dataset*

The IS statistics for nine case studies from the SPC/NSSL 2005 Spring Program dataset (Kain et al. 2008; Ahijevych et al. 2009, section 4) are aggregated as described in section 2c. For each of the IS statistics, 95% confidence intervals are evaluated by bootstrapping, in order to quantify the uncertainty associated with their variability within the cases. The 24-h lead-time precipitation forecasts produced by three models [2- and 4-km simulations of the NCAR version of the Weather Research and Forecast model (WRF2 and WRF4 NCAR) and 4-km simulations of the NCEP version of the WRF model (WRF4 NCEP)] are verified against the stage II radar-based analysis (STG2). All of the 1-h accumulation precipitation fields were remapped onto the stage II 4-km resolution grid. As described in section 2d, all cases are initially padded. Cropping, interpolating, and tiling are then performed, and the aggregated IS statistics are evaluated.

Figure 17 shows the aggregated IS skill score obtained by tiling, for the nine spring 2005 case studies, for the WRF2 and WRF4 NCAR and WRF4 NCEP models. All of the models exhibit similar behavior: small scales and large thresholds exhibit the worse skill, large scales exhibit positive skill, and the no-skill to skill transition scale corresponds to 32 grid points (128 km). Skill is usually not expected at the model grid resolution, or even up to 2–4 times this resolution. However, this large no-skill to skill transition scale indicates that all three models represent frontal features well (features larger than 100 km), but perform poorly for convective events (features smaller than 100 km). Small scales and low

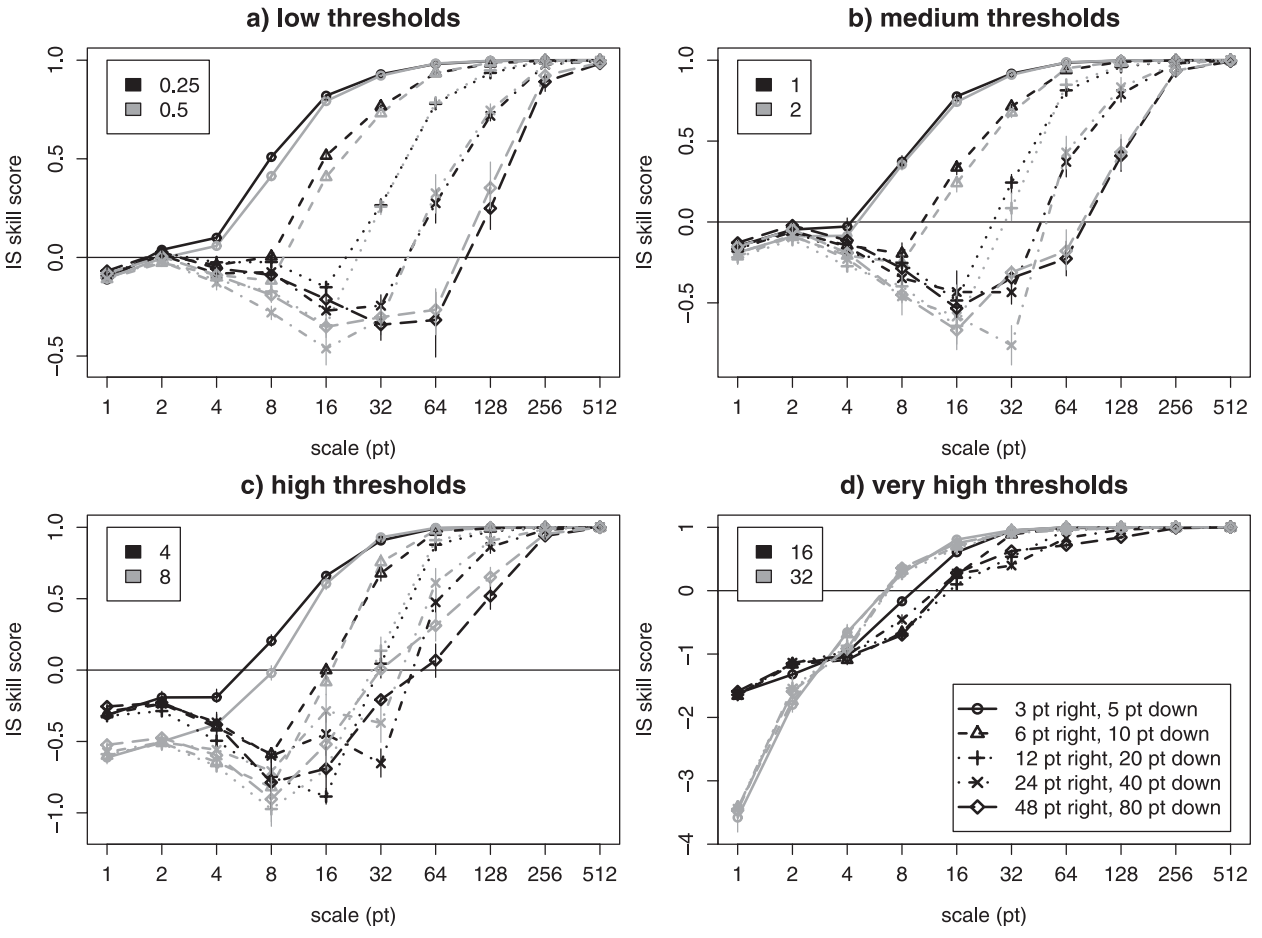

FIG. 13. Intensity-scale skill score obtained by tiling for the synthetically perturbed case with displacement error. The statistics for the different cases are plotted with different symbols and lines, as indicated in the bottom-right legend in (d). Different gray shades correspond to different precipitation thresholds (mm), as indicated in the top-left legend in each panel. Segments indicate the bootstrap 95% confidence intervals for the aggregated tiles.

thresholds exhibit slightly positive skill due, as for the geometric cases, to a small error associated with few events. The sole significant difference between the models is the poorer skill of the WRF4 NCEP model for high thresholds at the smallest scale (i.e., the 4-km model resolution scale). However, when considering increasing scales, already at resolutions equal to 2 or 4 times the model resolution, the skill levels of the three models are no longer significantly different. The behavior and differences between the three models shown by the cropped and interpolated cases are similar and are, therefore, not shown.

To illustrate the sensitivity of the IS statistics to the approach chosen to tackle the dyadic domain constraints, the IS skill scores obtained for the aggregated spring 2005 cases by cropping, interpolating, and tiling are compared. Figure 18 shows the aggregated IS skill scores for the WRF4 NCAR model. Results obtained for the WRF2 and WRF4 NCEP models are similar and are therefore not shown. When comparing cropping versus interpolating (cf. Figs. 18a and 18b) and cropping versus tiling (cf. Figs. 18a and 18c), for small thresholds and medium to large scales (in correspondence to the no-skill to skill transition scale), the cropped cases exhibit less skill. This is due to the removal of large areas of zeros and small values, by cropping both the forecast and the analysis, which correspond to correct rejections. When comparing interpolating versus cropping (cf. Figs. 18a and 18b) and interpolating versus tiling (cf. Figs. 18b and 18c), the skill levels on small scales are affected, since the nearest-neighbor interpolation removes isolated columns from the forecast and analysis fields, thus affecting the smallest scales. The interpolated cases on these small scales exhibit less skill than the cropped and tiled cases; however, such a loss of skill is purely due to random chance, since the effects of the

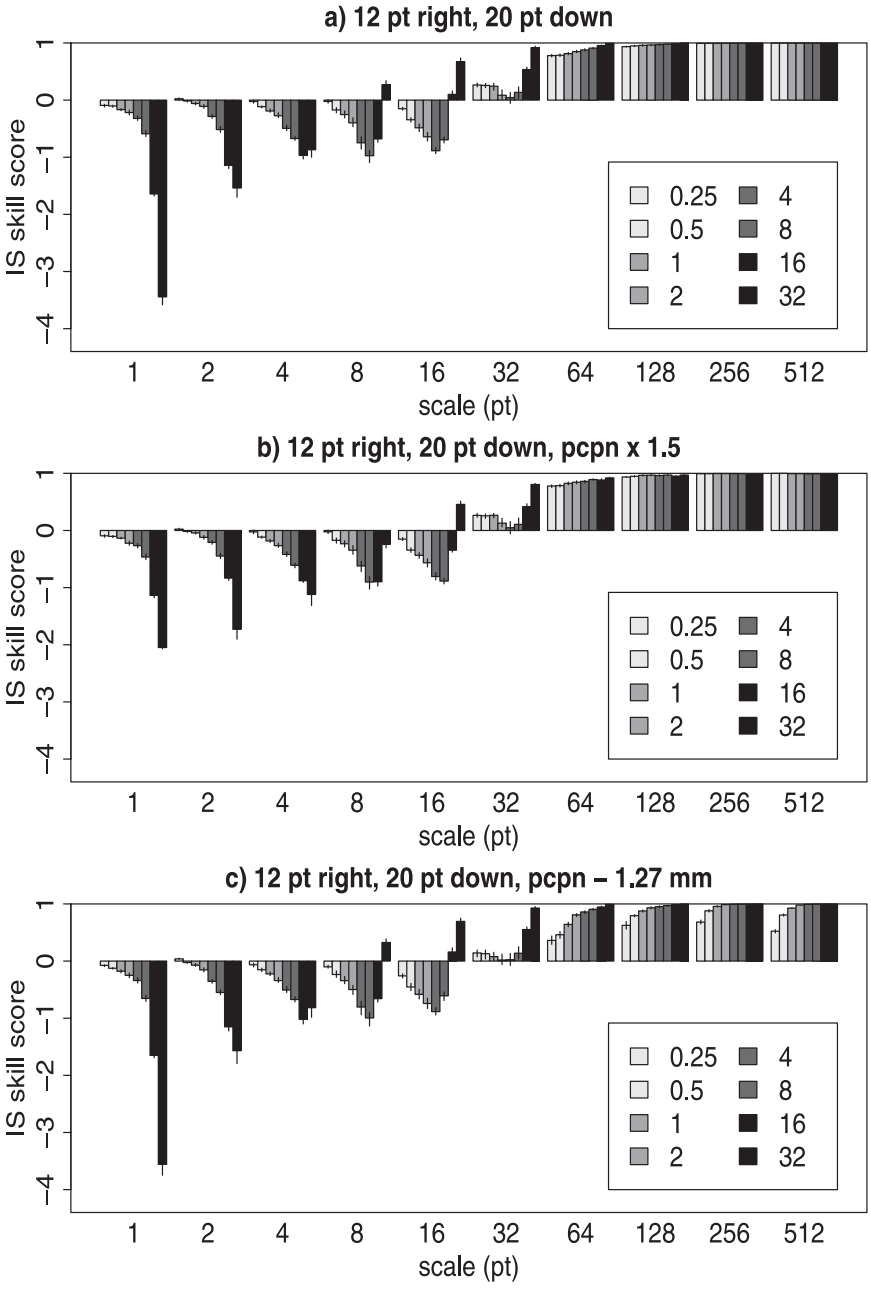

FIG. 14. Intensity-scale skill score obtained by tiling for the synthetically perturbed case with bias error. Bars of different gray shades correspond to increasing precipitation thresholds (mm), as indicated in the legend. Segments at the bar extremities indicate the bootstrap 95% confidence intervals for the aggregated tiles.

nearest-neighbor interpolation depend on which values are removed by the interpolation process itself. On the other hand, if a linear interpolation to a denser grid would be applied, one would expect to slightly increase the skill for small scales. In fact, such an interpolation would reduce the variability of the forecast and observation fields, and therefore the small-scale error would be reduced. The IS skill score produced by tiling seems in this case to be the most robust and reliable, since tiling does not involve any change of the original precipitation field.

The energy and energy relative difference are evaluated in order to assess the bias intensity-scale dependency for the spring 2005 dataset. Figure 19 shows the energy

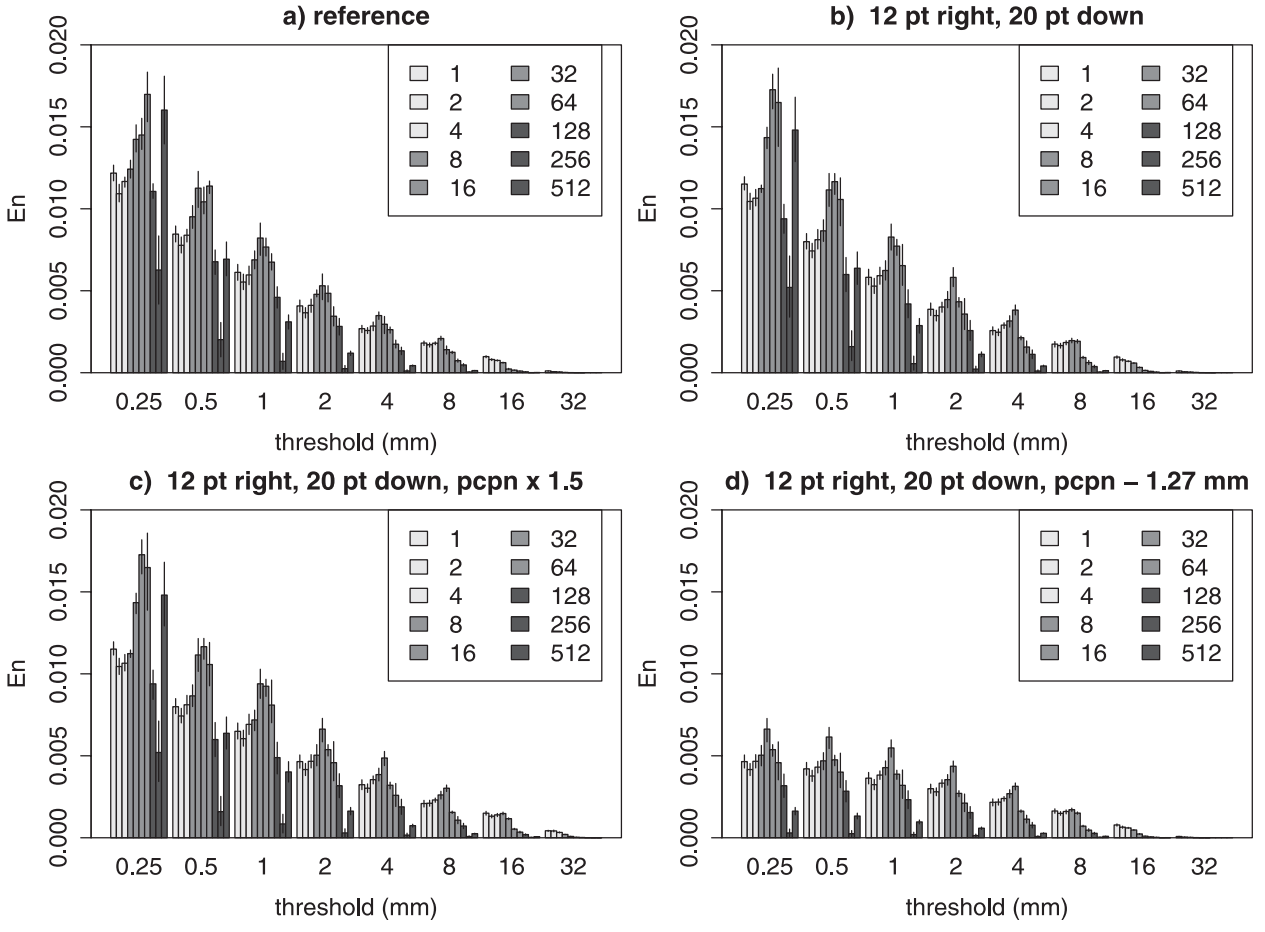

FIG. 15. Energy obtained by tiling for the synthetically perturbed case with bias error. Bars of different gray shades correspond to different scales (in grid points), as indicated in the legend. Segments at the bar extremities indicate the bootstrap 95% confidence intervals for the aggregated tiles.

obtained by tiling, for the stage II analysis and for the WRF2 and WRF4 NCAR and WRF4 NCEP models, for the aggregated case studies. The energy behavior pattern is similar to that for the synthetically perturbed case (section 3b; see Fig. 15): small thresholds are associated with larger energies (many events exceed the threshold) and large thresholds are associated with less energy. Scales of 8–16–32–64 grid points exhibit the largest energies, indicating that features ranging from 32 to 256 km characterize the fields. Figure 20 shows the energy relative difference obtained by tiling for the WRF2 and WRF4 NCAR and WRF4 NCEP models versus the stage II analysis, for the aggregated spring 2005 cases. All of the models tend to overforecast, especially for thresholds between 2 and 16 mm. However, the WRF4 NCEP model overforecasts more than the other two models, for all thresholds and scales. The WRF2 NCAR model underforecasts intense events, whereas the WRF4 NCAR model reproduces the intense events well. The comparison of the three models performed by analyzing the energy and energy relative difference for the cropped and interpolated cases led to similar results and is, therefore, not shown.

The sensitivity of the energy to the strategy chosen to tackle the dyadic domain constraints is also analyzed. Figure 21 shows the energy relative difference for the aggregated spring 2005 case studies, while comparing cropped versus interpolated cases, interpolated versus tiled cases, and tiled versus cropped cases. The statistics are illustrated for the WRF4 NCAR model. Results obtained for the WRF2 NCAR and WRF4 NCEP models are similar and are therefore not shown. When comparing cropping versus interpolating (Fig. 21a), the energy of the cropped fields is larger than those interpolated (especially at large scales), since cropping removes large regions with zeros and small values, whereas interpolating removes zeros and nonzeros (i.e., also large) values. When comparing interpolating versus tiling (Fig. 21b), the tiled fields exhibit larger energy

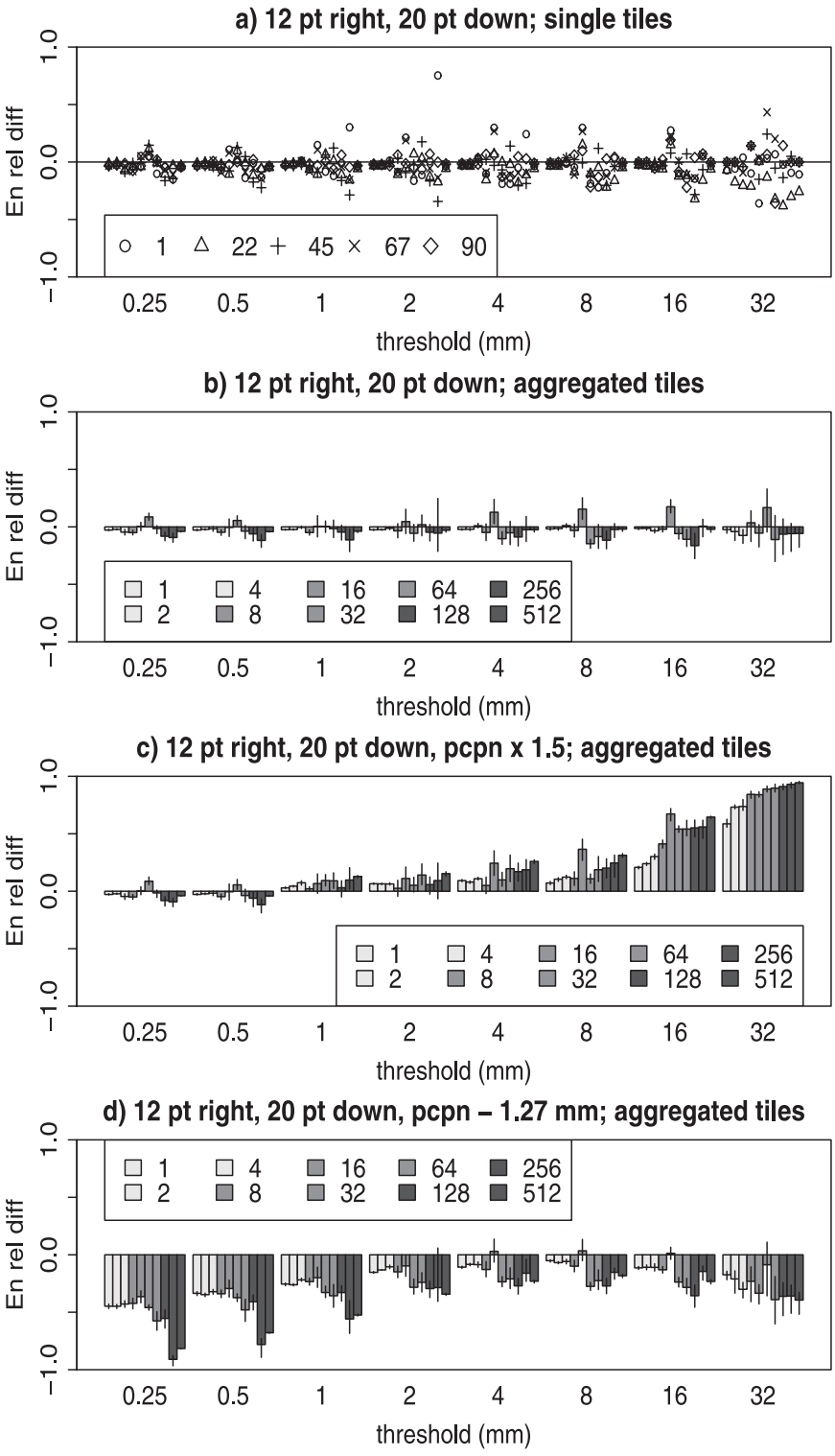

FIG. 16. Energy relative difference obtained for (a) each single tile and (b)–(d) by tiling, for the synthetically perturbed case with bias error. Bars of different gray shades correspond to different scales (in grid points), as indicated in the legend. Segments at the bar extremities indicate the bootstrap 95% confidence intervals for the aggregated tiles. Symbols in (a) correspond to the values of the energy relative difference for the five tiles, with tile origins indicated in the legend.

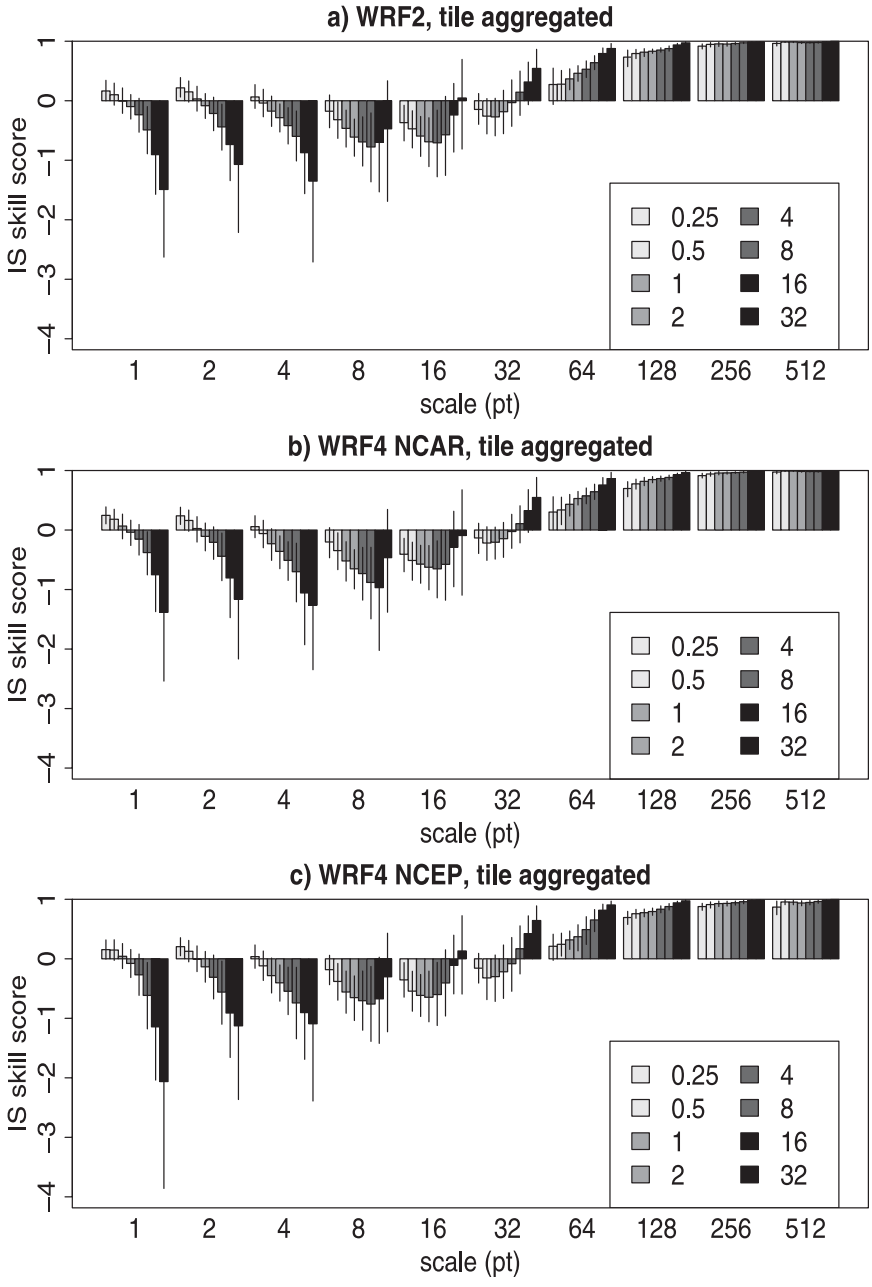

FIG. 17. Aggregated IS skill score obtained by tiling the domain for (a) WRF2 NCAR, (b) WRF4 NCAR, and (c) WRF4 NCEP for the spring 2005 cases. Bars of different gray shades correspond to increasing precipitation thresholds (mm), as indicated in the legend. Segments at the bar extremities indicate the bootstrap 95% confidence intervals for the aggregated cases.

than the interpolated fields; this is possibly due to an oversampling of the center of the domain, where the precipitation features are mainly concentrated, performed by the tiling procedure. When comparing tiling versus cropping (Fig. 21c), the energy of the cropped fields is slightly larger than the tiled fields; this shows that the removal of zeros and small values performed by the cropping has a larger effect than the oversampling performed by the tiling. These results, however, are not significantly different. The energy evaluated for the interpolated fields seems in this case to be the most reliable, since the nearest-neighbor interpolation randomly removes zero and nonzero precipitation values and, thus, does not artificially enhance the energy.

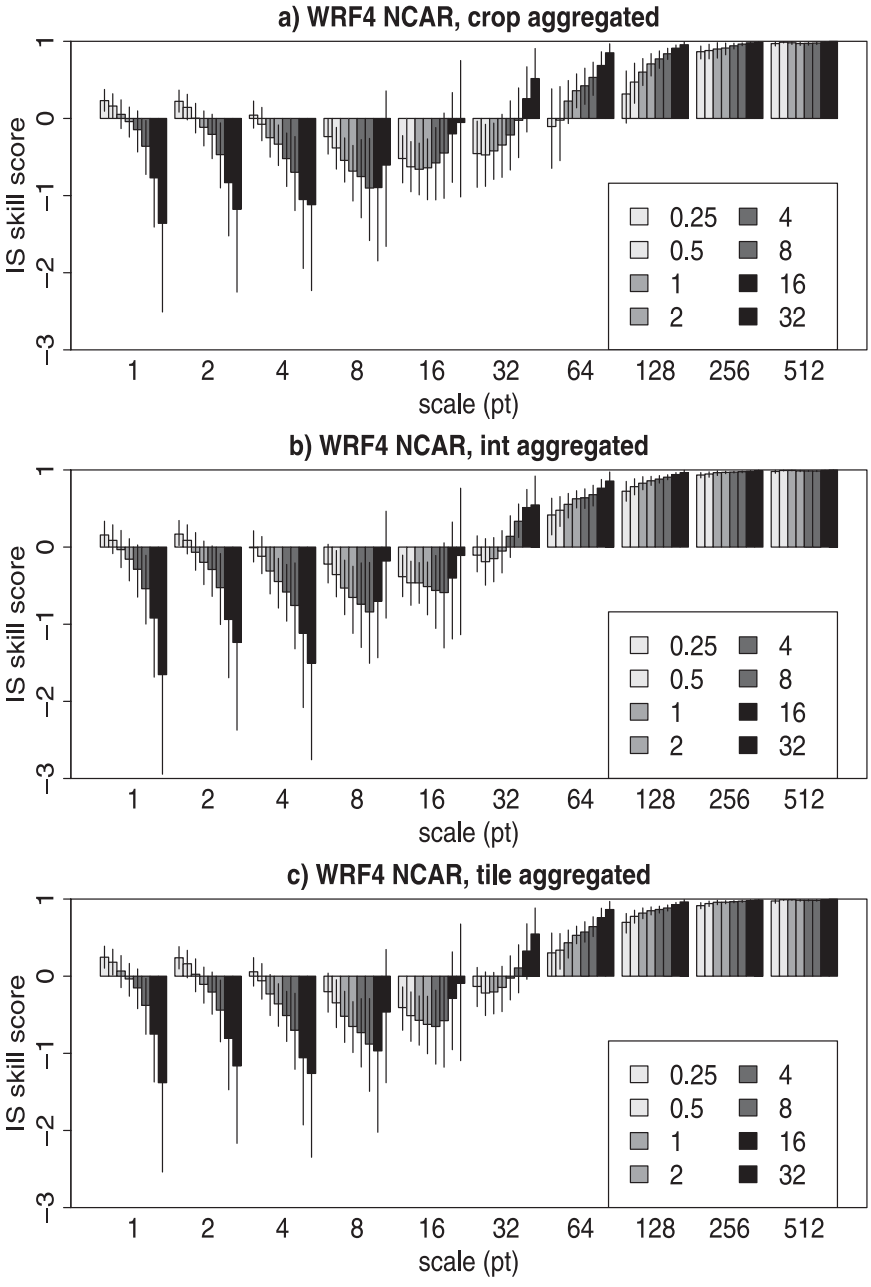

FIG. 18. Aggregated IS skill score for the WRF4 NCAR model obtained by (a) cropping the domain, (b) interpolating the values to a dyadic domain, and (c) tiling the domain for the spring 2005 cases. Bars of different gray shades correspond to increasing precipitation thresholds (mm), as indicated in the legend. Segments at the bar extremities indicate the bootstrap 95% confidence intervals for the aggregated cases.

The energy percentages and scale structures for the three models are similar (not shown). As is already evident from the energy relative difference (Fig. 20) for small thresholds (0.25–2 mm), all three models underforecast features at the scale of 8–32 grid points. These underforecast features correspond to round features (most probably caused by spurious radar echoes) visible in the stage II analysis but not present in any of the forecasts. For larger thresholds (4–16 mm), all three models exhibit underforecasting of small scales and a corresponding overforecasting of large scales (less so for the WRF4 NCEP model), which may be due to

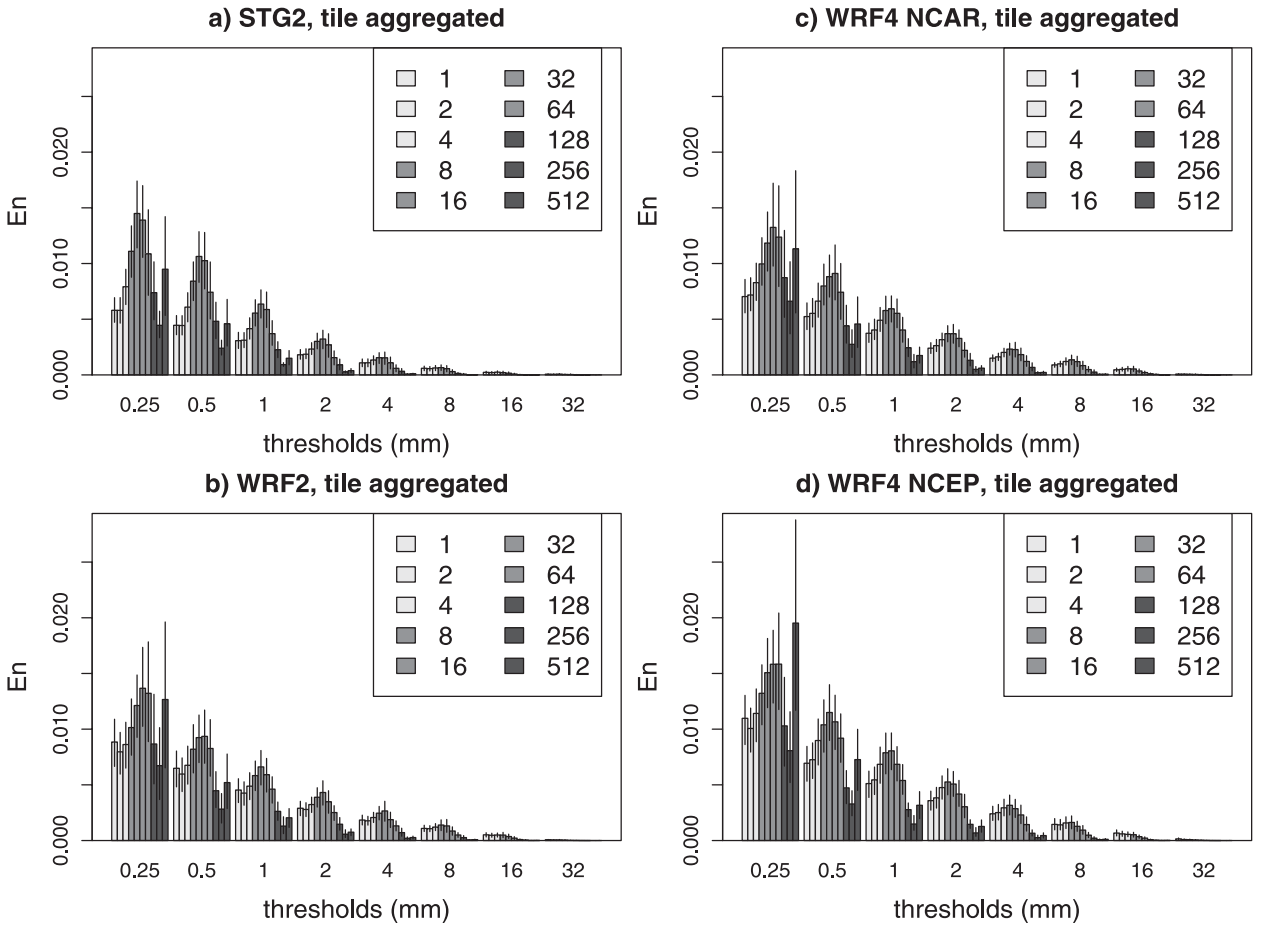

FIG. 19. Energy obtained by tiling the domain for (a) STG2, (b) WRF2 and (c) WRF4 NCAR, and (d) WRF4 NCEP for the aggregated spring 2005 cases. Bars of different gray shades correspond to different scales (in grid points), as indicated in the legend. Segments at the bar extremities indicate the bootstrap 95% confidence intervals for the aggregated cases.

a smoothing of the precipitation fields, which is commonly performed by NWP systems.

4. Conclusions

The IS technique introduced by Casati et al. (2004) is revisited and improved. Recalibration, which was formerly applied to separate bias from skill assessment, is no longer performed and the IS skill score for biased forecasts is evaluated. Energy and the proportion of energy at each spatial scale are then introduced to assess the bias on different scales and the spatial scale structure of the precipitation fields. Note that energy and energy percentages inform the user about the differences between the forecast and observation marginal distributions, whereas the IS skill score pertains to their joint distribution. Aggregation of the IS statistics for multiple cases is performed and confidence intervals are provided by bootstrapping. The bootstrapped samples were compiled from different cases and different dyadic tile positions.

The IS verification has been applied to the Intercomparison of Spatial Forecast Verification Methods dataset. The IS skill score assesses the skill for different precipitation intensities and on different spatial scales, separately. The geometric cases show that the spatial scales of the error are attributed to both the size of the features and their displacement. The geometric and synthetically perturbed cases show that the IS verification statistics are sensitive to displacements and bias errors. Moreover, precipitation fields are characterized by the presence of physically coherent features: intense precipitation events are in general characterized by small scales (e.g., convection), whereas large-scale events are usually associated with stratiform precipitation and moderate values. The NIMROD and synthetically perturbed cases show that this intrinsic relationship existing between the feature intensity and

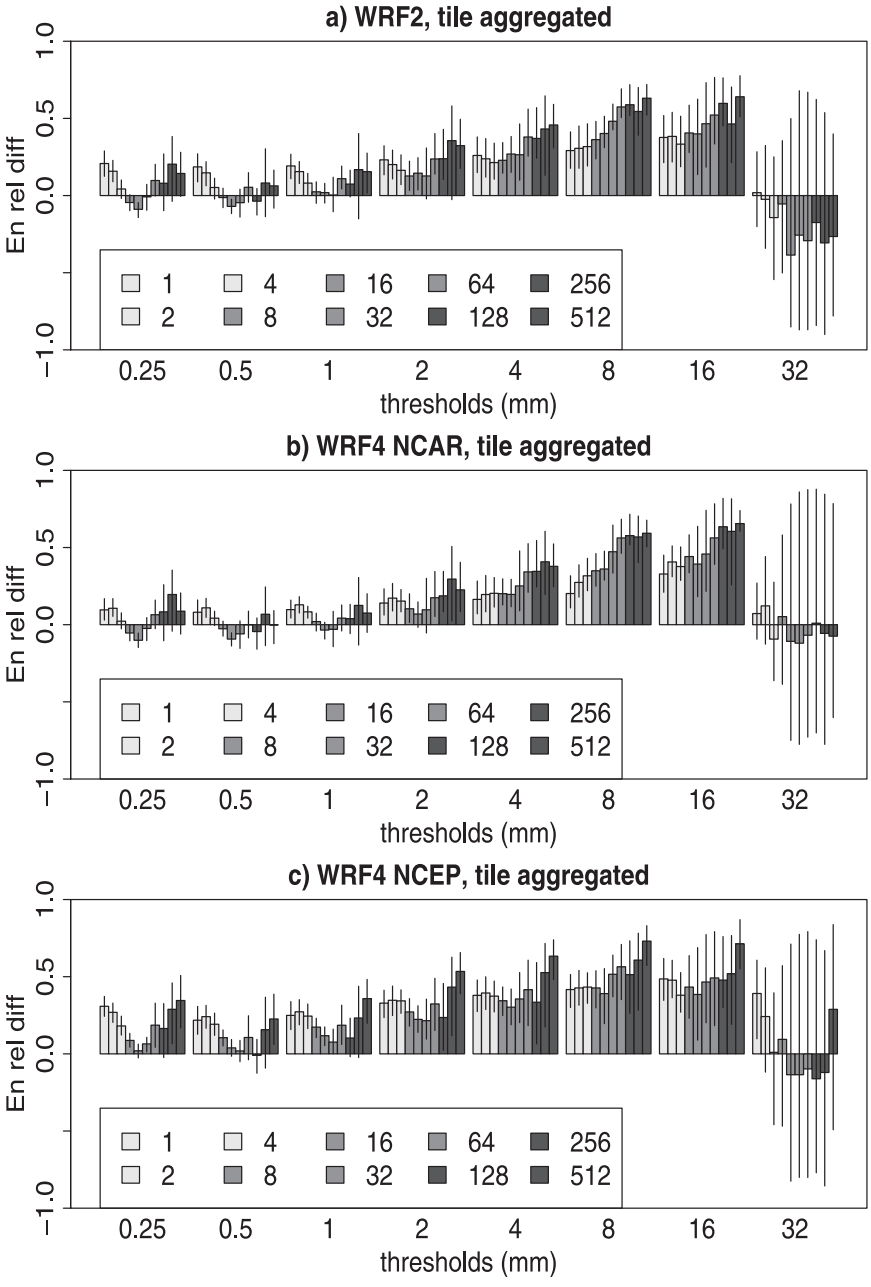

FIG. 20. Energy relative difference obtained by tiling the domain for (a) WRF2 and (b) WRF4 NCAR and (c) WRF4 NCEP for the aggregated spring 2005 cases. Bars of different gray shades correspond to different scales (in grid points), as indicated in the legend. Segments at the bar extremities indicate the bootstrap 95% confidence intervals for the aggregated cases.

scale is well captured by the IS statistics. The energy percentages enable the user to objectively analyze the scale structure of the fields and the scale-intensity relationship due to the spatial coherence of the precipitation features.

The IS skill score associated with the geometric cases enables us to highlight the differences between scale-separation and neighborhood verification techniques. In fact, on small scales the IS skill score is positive (no small-scale events are present and, therefore, little error is associated with these scales), whereas for neighborhood techniques (e.g., see Mittermaier and Roberts 2010) the skill on small scales is negative (the neighborhood size is too small to embed and smooth out the

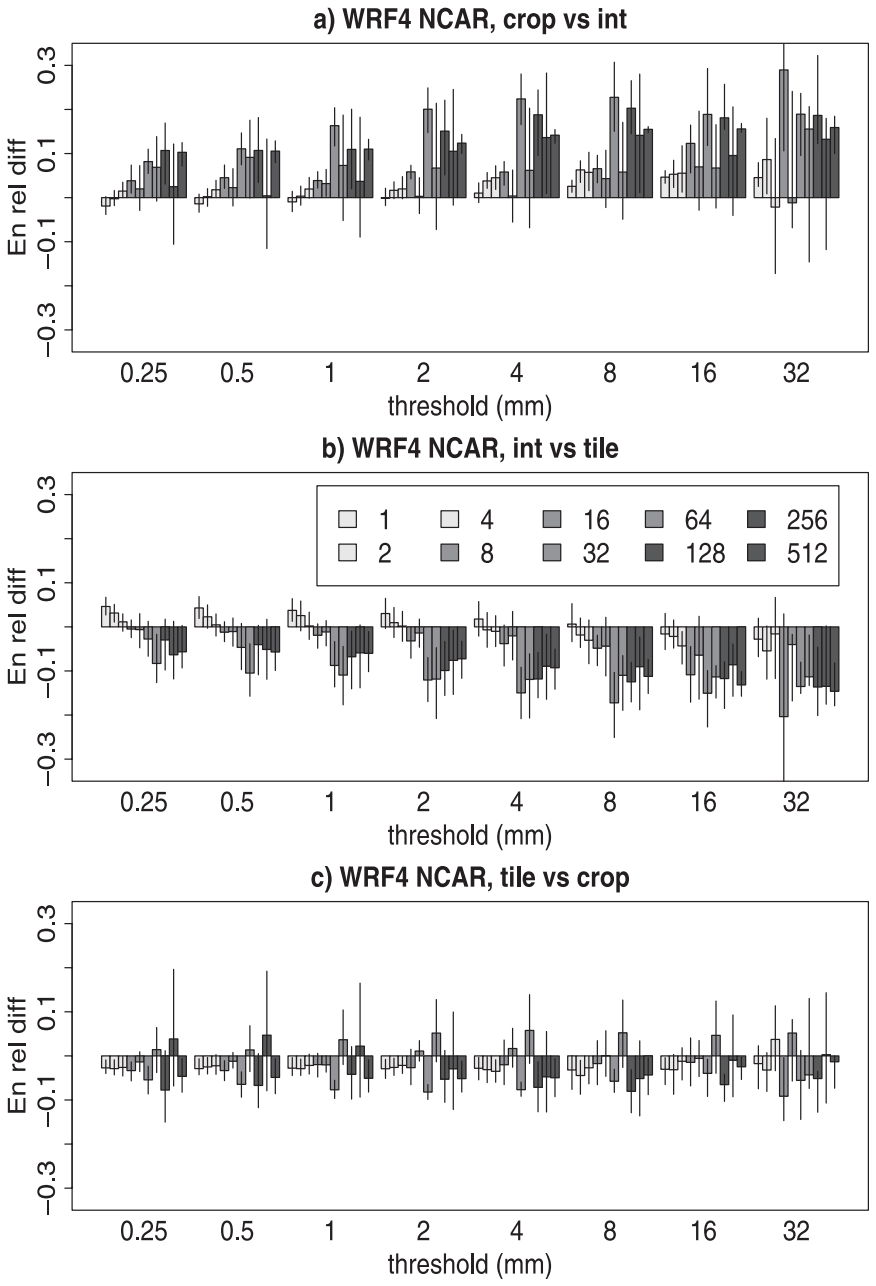

FIG. 21. Energy relative difference for the WRF4 NCAR model when comparing (a) cropping vs interpolation, (b) interpolation vs tiling, and (c) tiling vs cropping for the aggregated spring 2005 cases. Bars of different gray shades correspond to different scales (in grid points), as indicated in the legend. Segments at the bar extremities indicate the bootstrap 95% confidence intervals for the aggregated cases.

displacement of the ellipses). This follows from the different definition of ''scale'' for the scale-separation and neighborhood approaches. The IS scale components are defined by a single-band filter: the IS approach is therefore capable of isolating scales associated with different wavelengths, and of assessing the forecast error and skill for each individual scale, separately. On the other hand, for neighborhood verification techniques the scales are defined by a low-bandpass filter (smoothing): as the size of the neighborhood (scale) is increased, the forecast and observations are subject to a filtering process where the exact space–time matching requirements become more and more relaxed. Neighborhood verification approaches, therefore, by definition exhibit larger

skill for increasing scales; these approaches do not aim to separate the scales but assess the neighborhood size (or critical smoothing scale) above which skill is achieved. Scale-separation and neighborhood verification approaches provide different information about the scale dependency of the forecast skill (see also Gilleland et al. 2009; Ebert 2009; Casati et al. 2008).

Aggregated IS skill scores for the SPC/NSSL 2005 Spring Program dataset do not show significant differences in the skill levels of the three models assessed: WRF2 and WFR4 NCAR and WRF4 NCEP. All three models exhibit positive skill on large scales, with the worst skill found on small scales and at large thresholds. The no-skill to skill transition scale occurs at the 32-gridpoint (128 km) scale, indicating good representation of synoptic scales but poor performance for convective precipitation. The energy bias reveals that all three models overforecast medium to large scales, and that the WRF4 NCEP model overforecasts more remarkably than the other two models, for all scales and thresholds. To better visualize the differences between the NWP performance, the differences in the IS verification statistics for the different models can be evaluated, and their significance can be assessed by bootstrapping (not shown).

The systematic pattern of the IS skill score obtained for the spring 2005 dataset, showing negative skill for small scales and positive skill for large scales, is quite typical. Note that this pattern is associated with the typical characteristics of gridded quantitative precipitation forecasts; in fact, precipitation forecasts usually represent well large-scale weather phenomena (e.g., frontal systems), but perform poorly for the less predictable, intense small-scale events, and tend to be noisy and nonskillful at the forecast resolution scale. Note also that for the neighborhood verification approaches this systematic pattern (i.e., larger skill with increasing smoothing scale) is expected, whereas for the IS skill score this is not the case. In fact, the IS skill score can also exhibit different patterns of behavior. As an example, the geometric cases have positive skill for small scales, due to the small error associated with these scales, and the skill becomes negative for the larger scales, associated with the feature size and displacement errors. Moreover, the IS skill score is capable of identifying specific scale-dependent errors related to individual cases; as an example, the displacement of the storm for the NIMROD case study is detected by the negative skill at the 160-km scale, whereas the no-skill to skill transition scale for the NIMROD forecasts is 40 km (Casati et al. 2004). When performing a verification of several forecasts, it is often desirable to identify individual forecasts that stand out because of a particularly good or poor performance; this can be done, as an example, by evaluating the difference between the IS skill score for individual forecasts versus the aggregated IS skill score. As an alternative, a variation of the IS skill score could be defined, where the $\mathrm{MSE}_{u,l}$ for the single case is compared against the $\mathrm{MSE}_{u,l}$ aggregated for all forecasts.

The sensitivity of the IS statistics to the discrete nature of the wavelet support has been analyzed. Tiling the forecast domain with randomly displaced dyadic supports and aggregating the IS statistics associated with each tile enables the user to virtually eliminate the effects due to the discrete wavelet support. The optimal number of tiles to be used depends on the forecast spatial characteristics and can be determined by sensitivity tests. In this article, such an optimal number is estimated by the minimum number of tiles needed to obtain energy relative differences not significantly different from zero, for forecasts and observations that are expected to have identical energies. For precipitation fields, a small number of tiles is sufficient to eliminate the wavelet support effects, whereas smoother fields need more tiling. This is due to the characteristic square shape of the Haar wavelets, which efficiently represents highly discontinuous and noisy fields, such as precipitation. When aggregating cases from multiple model runs, the tiling constraints are relaxed because the weather moves across the spatial domain and features assume naturally different positions with respect to the discrete wavelet support.

Four different approaches addressing the dyadic domain constraint have been discussed: padding, cropping, interpolating, and tiling. These approaches were compared using the results obtained for the spring 2005 dataset aggregated statistics. Padding (cropping) can enhance (diminish) the IS skill score on medium to large scales and small intensities because of the addition (removal) of correct zeros (and small values). Interpolation changes the original field values and affects the IS statistics on the smallest scales. Tiling can provide slightly misrepresentative IS verification statistics, because of the oversampling of the center of the domain. The differences between the results, however, are very marginal, and the overall qualitative behavior of the IS statistics does not change significantly depending on the approach chosen. When aggregating multiple cases, the choice of the approach to be used to tackle the dyadic domain constraints should be based on the forecast characteristics and verification purposes. On the other hand, for single case studies, tiling provides the most robust and reliable approach, since it smoothes the effects due to the discrete wavelet support and is not affected by any change in the original precipitation fields.

An alternative approach to eliminate the effects due to the position of the discrete wavelet support with respect to the precipitation features could be to use continuous wavelet transforms (see van den Berg 2004, chapter 2 and references therein) rather than discrete wavelet transforms and tiling. In fact, continuous wavelet transforms use continuously varying translation and dilatation parameters; they are therefore shift invariant and they diagnose the spectral components of the transformed fields for continuously varying scales. Continuous wavelets would then eliminate the dyadic support effects and enrich the diagnostic power of the IS statistics, providing a continuous spectrum of scales. However continuous wavelet transforms produce a redundant representation of the transformed field, so that the orthogonality of the scale components and additive properties of the IS statistics (and their percentages) would be lost. Future work could investigate such an alternative approach, its characteristics, and implications.

The Intercomparison of Spatial Forecast Verification Methods (Gilleland et al. 2009) has provided a great opportunity to further develop the IS verification approach and to respond to some users' needs. Moreover, an new open source code for the evaluation of the IS verification statistics is now available within the Meteorological Evaluation Toolkit (MET; available online at http://www.dtcenter.org/met/users/), along with the one already available in the verification package developed by M. Pocernich (available online at http://cran.r-project.org/web/packages/verification/) for the R statistical language (http://www.R-project.org). A more widespread use of the new spatial verification approaches is encouraged and can help the verification approach developers to better address user-relevant issues. More generally, the intercomparison project has provided a common framework to enhance our understanding and to compare the capabilities of the new spatial verification methods, which benefits both technique developers and potential users. Metaverification intercomparisons provide guidance for the users in selecting the appropriate technique that will provide specific information on the forecast quality. Communication between the developers and the user community is fundamental for a more meaningful verification.

Acknowledgments. This work was started during my visit at NCAR, under the DTC/WRF visitor program. I thank B. Brown, E. Gilleland, D. Ahijevych, E. Ebert, M. Mittermaier, and L. Lefaivre for encouraging me to participate in the intercomparison project and for the very stimulating discussions. I thank J. Halley-Gotway for coding the IS technique in the Meteorological Evaluation toolkit. Finally, I thank the three reviewers, and in particular reviewer C, for their comments which led to substantial improvements to the article.

REFERENCES

Ahijevych, D., E. Gilleland, B. G. Brown, and E. E. Ebert, 2009: Application of spatial verification methods to idealized and NWP gridded precipitation forecasts. *Wea. Forecasting,* **24,** 1485–1497.

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

Burt, P. J., and E. H. Adelson, 1983: The Laplacian pyramid as a compact image code. *IEEE Trans. Commun.,* **31,** 532–540.

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

——, and Coauthors, 2008: Forecast verification: Current status and future directions. *Meteor. Appl.,* **15,** 3–18.

Cisma, G., and A. Ghelli, 2008: On the use of the intensity-scale verification technique to assess operational precipitation forecasts. *Meteor. Appl.,* **15,** 145–154.

Daubechies, I., 1992: *Ten Lectures on Wavelets*. SIAM, 357 pp.

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

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

——, 2009: Neighborhood verification: A strategy for rewarding close forecasts. *Wea. Forecasting,* **24,** 1498–1510.

——, and J. L. McBride, 2000: Verification of precipitation in weather systems: Determination of systematic errors. *J. Hydrol.,* **239,** 179–202.

Efron, B., and R. J. Tibshirani, 1993: *An Introduction to the Bootstrap*. Chapman and Hall, 436 pp.

Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of spatial forecast verification methods. *Wea. Forecasting,* **24,** 1416–1430.

Golding, B. W., 2000: Quantitative precipitation forecasting in the UK. *J. Hydrol.,* **239,** 286–305.

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

Hoffman, R., 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, 240 pp.

Kain, J. S., and Coauthors, 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Wea. Forecasting,* **23,** 931–952.

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.

Mallat, S. G., 1989: A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.,* **11,** 674–693.

Mittermaier, M. P., 2006: Using an intensity-scale technique to assess the added benefit of high-resolution model precipitation forecasts. *Atmos. Sci. Lett.,* **7,** 36–42.

——, and N. Roberts, 2010: Intercomparison of spatial forecast verification methods: Identifying skillful spatial scales using the fractions skill score. *Wea. Forecasting,* **25,** 343–354.

Murphy, A. H., and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.,* **115,** 1330–1338.

Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.,* **136,** 78–97.

van den Berg, C. J. Ed., 2004: *Wavelets in Physics*. Cambridge University Press, 478 pp.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 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. Geophys. Res.,* **105,** 10 129–10 146.