Spatial Forecast Verification Methods Intercomparison Project: Application of the SAL Technique

HEINI WERNLI, CHRISTIANE HOFMANN, AND MATTHIAS ZIMMER
Institute for Atmospheric Physics, University of Mainz, Mainz, Germany

(Manuscript received 6 February 2009, in final form 4 May 2009)

ABSTRACT

In this study, a recently introduced feature-based quality measure called SAL, which provides information about the structure, amplitude, and location of a quantitative precipitation forecast (QPF) in a prespecified domain, is applied to different sets of synthetic and realistic QPFs in the United States. The focus is on a detailed discussion of selected cases and on the comparison of the verification results obtained with SAL and some classical gridpoint-based error measures. For simple geometric precipitation objects it is shown that SAL adequately captures errors in the size and location of the objects, however, not in their orientation. The artificially modified (so-called fake) cases illustrate that SAL has the potential to distinguish between forecasts where intense precipitation objects are either isolated or embedded in a larger-scale low-intensity precipitation area. The real cases highlight that a quality assessment with SAL can lead to contrasting results compared to the application of classical error measures and that, overall, SAL provides useful guidance for identifying the specific shortcomings of a particular QPF. It is also discussed that verification results with SAL and other error measures should be interpreted with care if considering large domains, which may contain meteorologically distinct precipitation systems.

1. Introduction

During the past few years, many novel techniques have been developed to assess the quality of quantitative precipitation forecasts (QPFs). A major aim of these methods is to overcome the double-penalty problem, which is inherent in classical gridpoint-based verification strategies (Jolliffe and Stephenson 2003) when applied to richly structured QPFs. This problem arises for instance if an observed precipitation feature is displaced in the forecast. An error measure like, for instance, the mean error punishes this forecast for missing the actual feature and for predicting it in a region where no precipitation has occurred. However, an alternative less useful forecast that totally misses the event is only punished once. In addition to better coping with this double-penalty issue, the new verification techniques try to provide useful information about the specific character of the forecast error. An overview on and a categorization of these methods can be found in the recent reviews by Casati et al. (2008), Rossa et al. (2008), Ebert (2008), and Gilleland et al. (2009).

Wernli et al. (2008, subsequently referred to as WPHF) proposed the three-component feature-based quality measure SAL, which considers the structure (S), amplitude (A), and location (L) of a QPF. Application of the technique requires the preselection of a domain of interest (e.g., a river catchment), the choice of a threshold contour value for independently identifying precipitation objects in the observational data and model forecast (which should be available on the identical grid), and eventually the calculation of the three components of SAL according to the equations given in section 2 of WPHF (see also section 2a below). So far, SAL has been applied to synthetic precipitation fields and a large set of operational QPFs from the nonhydrostatic regional model from the Consortium for Small-Scale Modeling (COSMO) with a horizontal resolution of 7 km for the Elbe catchment (WPHF) and for other catchments in Germany (Paulat 2007). Hofmann et al. (2009) used SAL to identify poor and excellent COSMO model QPFs during the years 2002–06. SAL has also been utilized to assess the quality of QPFs in Switzerland (Jenkner 2008) and to systematically compare the QPF performance of 20 regional models in southern Germany during summer...
2007 (Zimmer et al. 2008; Zimmer and Wernli 2008). In addition, the technique has been used to compare the quality of precipitation estimates derived from satellites and a regional numerical model (Früh et al. 2007).

In this study, SAL is applied to QPFs in subregions of the United States, mainly in the Midwest and parts of southern United States. These QPFs form the basis of the spatial forecast verification methods intercomparison (Ahijevych et al. 2009, AGBE). The aim is to test the suitability of SAL for the selected intercomparison cases and to compare the results obtained by SAL with those from classical gridpoint-based error measures, that is, the mean error (ME), the root-mean-square error (RMSE), the frequency bias (FBI), and the Heidke skill score (HSS). Section 3 presents the geometric cases, section 4 the artificially modified QPFs, and section 5 some real forecasts. A central goal of the study is to elucidate the situations when SAL and classical error measures provide contrasting guidance about the quality of QPFs, as summarized in section 6. But first, section 2 provides a brief overview on SAL and the classical error measures.

2. SAL and classical error measures

a. The quality measure SAL

As outlined in detail by WPHF, SAL is a three-component feature-based quality measure for QPFs, where the three components aim at quantifying the quality of the forecast in terms of its structure (S), amplitude (A), and location (L) in a prespecified region of interest, for instance, a major river catchment. For the structure and location components, it is necessary to identify coherent objects in the observed and predicted precipitation fields; however, a one-to-one matching between the identified objects in the observed and simulated precipitation fields is not required. This constitutes a major difference of SAL compared to, for instance, the “contiguous rain area” (CRA) method of Ebert and McBride (2000) and the object-based technique introduced by Davis et al. (2006), which has been further developed and is now referred to as the Method for Object-Based Diagnostic Evaluation (MODE) approach (Davis et al. 2009). The practical application of SAL involves the following three steps:

1) The specification of a domain (i.e., a river catchment, a country or state, or a numerical model domain)—In WPHF, examples for the application of SAL have been shown for the German part of the catchment of the river Elbe. Here, subjectively defined rectangular domains will be chosen as described below.

2) The choice of a threshold to define precipitation objects—In WPHF, this threshold value has been taken as \( R^* = \frac{1}{15} R_{\text{max}} \), where \( R_{\text{max}} \) denotes the maximum gridpoint value of the precipitation that occurs within the considered domain. Since this maximum value is sensitive to outliers (e.g., single grid points with very intense precipitation), the 95th percentile of all gridpoint values in the domain larger than 0.1 mm (denoted as \( R^5_{\text{95}} \)) is used in this study (and in our later applications of SAL) instead of \( R_{\text{max}} \), that is,

\[
R^* = \frac{1}{15} R^5_{\text{95}}. \tag{1}
\]

3) The calculation of the three components of SAL is outlined briefly in the following paragraphs and in more detail in section 2 of WPHF.

The amplitude component A corresponds to the normalized difference of the domain-averaged precipitation values of the model \( R_{\text{mod}} \) and the observations \( R_{\text{obs}} \):

\[
A = \frac{D(R_{\text{mod}}) - D(R_{\text{obs}})}{0.5[D(R_{\text{mod}}) + D(R_{\text{obs}})]}. \tag{2}
\]

Here, \( D(R) \) denotes the domain average of the precipitation field \( R \). The values of \( A \) are within \([-2 \ldots +2]\) and 0 denotes a perfect forecast in terms of amplitude.

The location component of SAL consists of two parts, \( L = L_1 + L_2 \), where the first term measures the normalized distance between the centers of mass of the modeled and observed precipitation fields: \( L_1 = \frac{|x(R_{\text{mod}}) - x(R_{\text{obs}})|}{d} \)

where \( d \) denotes the largest distance between two boundary points of the considered domain and \( x(R) \) denotes the center of mass of the precipitation field \( R \) in this domain. The values of \( L_1 \) are in the range \([0 \ldots 1]\) and a value of \( L_1 = 0 \) indicates that the centers of mass of the predicted and observed precipitation fields are identical. The second part, \( L_2 \), measures the averaged distance between the center of mass of the total precipitation fields and the individual precipitation objects (see WPHF for the details). \( L_2 \) can only differ from zero if either the observations or the forecast (or both) contain more than one object in the considered domain. The scaling of \( L_2 \) is such that it has the same range as \( L_1 \), and hence the total location component \( L \) can reach values between 0 and 2.

Finally, for the structure component S the basic idea is to compare the volume of the normalized precipitation objects. As shown in WPHF, such a measure captures information about the size and shape of precipitation objects. For every object \( R_\alpha \), a “scaled volume” \( V_\alpha \) is
calculated as the following sum over all gridpoint values $R(i, j)$:

$$V_n = \sum_{(i,j) \in D} \frac{R(i, j)}{R_{n, \text{max}}},$$

(4)

where $R_{n, \text{max}}$ denotes the maximum precipitation value within the object. The volume $V_n$ is calculated separately for all objects in the observational and forecast datasets. Then, the weighted mean of all objects’ scaled precipitation volume, referred to as $S$, is defined as the normalized difference in $V$, analogously to the $A$ component:

$$S = \frac{V(R_{\text{mod}}) - V(R_{\text{obs}})}{0.5[V(R_{\text{mod}}) + V(R_{\text{obs}})]}.$$  

(5)

The scaling is again such that the possible range of values extends from $-2$ to $+2$. Positive values of $S$ indicate that the predicted precipitation objects are too large and/or too flat; in contrast, negative values occur for too small and/or too peaked objects.

b. Classical error measures

In addition to SAL, the following gridpoint-based error measures, here referred to as classical error measures, will be used in this study.

- The mean error is defined as the average difference between the observations and the forecasts at all grid points in the domain:

$$\text{ME} = \frac{1}{N} \sum_{(i,j) \in D} [R_{\text{mod}}(i, j) - R_{\text{obs}}(i, j)],$$

(6)

where $N$ is the number of grid points in the considered domain $D$. For perfect forecasts, the ME is equal to zero; positive and negative values (in mm) denote an over- or underestimation of the precipitation amounts, respectively. The amplitude component $A$ is related to the ME, since it corresponds to the ME scaled by the averaged observed and predicted precipitation amounts. Therefore, the relative measure $A$ and the absolute measure ME have always the same sign, but whereas a 10% overestimation of the precipitation amount in the considered domain leads always to the same value of $A = 0.095$, the ME will depend on the absolute precipitation values.

- The root-mean-square error, calculated as

$$\text{RMSE} = \left\{ \frac{1}{N} \sum_{(i,j) \in D} [R_{\text{mod}}(i, j) - R_{\text{obs}}(i, j)]^2 \right\}^{0.5},$$

(7)

is always positive (in mm). For a perfect forecast its value is zero. In addition, a normalized RMSE has been calculated as follows: $\text{nRMSE} = \text{RMSE}/D(R_{\text{obs}})$, where again $D(R_{\text{obs}})$ denotes the domain-averaged observed precipitation value. This normalized RMSE will be useful when comparing forecasts of light and intense precipitation intensity.

- The FBI for a low-gridpoint threshold value of 0.1 mm is calculated from the so-called contingency table; it corresponds to the ratio of the number of predicted to the number of observed events. Obviously, 1 corresponds to the optimum value and values larger (smaller) than 1 indicate a systematic overforecasting (underforecasting) of the event.

- Finally, the Heidke skill score is calculated for the same threshold of 0.1 mm. The HSS ranges from minus infinity to 1 (the optimum value), with 0 indicating a forecast of no skill. The HSS measures the fraction of correct forecasts after eliminating the forecasts that would be correct due purely to random chance.

For a more in-depth discussion of these error measures, the reader is referred to, for example, Jolliffe and Stephenson (2003).

3. Geometric cases

A set of synthetic forecasts has been produced for the intercomparison project (AGBE), which here are referred to as the “geom”-cases G0–G7 (G0 corresponds to geom000, etc.). They are all characterized by a single precipitation object with a simple geometric shape. Due to the particular shape of the objects [see Eq. (1) in AGBE], the structure component $S$ and the amplitude component $A$ are always identical. Therefore, the discussion will only consider the components $A$ and $L$.

Table 1 presents the values of these parameters and of the classical error measures for four pairs of geom-cases. As for the real-case forecasts presented below, the data have been interpolated onto a regular latitude–longitude grid with 0.1° resolution, which does not cover the entire domain of the original dataset.\footnote{This regridding was performed in order to make the data compatible with our graphics software. As a side effect, our values of the gridpoint-based error measures differ from the ones given, for instance, by AGBE, who performed the calculations on the original grid.}

The first comparison, G0 versus G2 (where G0 is assumed to be the true QPF and G2 the forecast; see Figs. 1a and 1b), corresponds to a displacement of the precipitation object in the forecast. This leads to a relatively large value of $L = 0.39$ and (almost) zero values of $A$.
The ME is close to zero because the positive and negative deviations cancel each other; that is, the ME is not able to identify the displacement. Also, the FBI value close to 1 does not indicate a problem with the forecast and the RMSE is (although clearly positive) the smallest of the four considered cases. In sharp contrast, the HSS is close to zero indicating that this forecast has no skill. This simple example already illustrates the diversity in the quality assessment when using classical error measures: ME and FBI treat the forecast as being close to perfect, the RMSE is not perfect but is still the best of the four measures, whereas the HSS indicates no skill. In contrast, SAL provides the more meaningful information: that the amplitude and structure of the QPF are (nearly) perfect, and that the predicted QPF is significantly displaced compared to the observed one.

For the comparisons of G0 versus G3 and G0 versus G5 (in both cases G0 is taken as the truth), the object in the forecast is both shifted and too large (Fig. 1). This is well captured by SAL (positive value of $L$ but smaller than for G0 versus G2; strongly positive values of $A$ and $S$). The classical measures all provide poor values, in particular for G0 versus G5, where the ME and RMSE are largest, the FBI is very large, and the HSS close to zero. The reason for this poor assessment of G0 versus G5 when using the classical parameters (compared with G0 versus G2) is related to the fact that classical measures like ME, RMSE, and FBI are strongly sensitive to the spatial extent of the error. The assessment by SAL is more to the point: G0 versus G2 is worse in terms of $L$, whereas G0 versus G3 and G0 versus G5 are (much) worse in terms of $A$ and $S$. Clearly, the preference for QPFs with low values of one of the three components varies for different users and applications.

The last comparison (G3 versus G5) is interesting because it is clearly the best in terms of the HSS and the location component of SAL. The other SAL parameters indicate that G5 overestimates the size and amount of precipitation, which goes along with intermediate values of the ME, RMSE, and FBI. It appears from these idealized examples that the HSS is particularly sensitive to displacement errors and that a significant overlap of the precipitation objects is required for a substantially positive HSS value.

These results are consistent with the analysis of another (more variable) set of idealized geometric QPFs by WPHF (their section !b).

4. Fake cases

This set of QPFs has been produced by artificially modifying the 1 June 2005 real-case simulation performed with the 2-km-resolution version of the Weather Research and Forecasting (WRF) model (see section 5). In this section, we compare three of these so-called fake cases with the undisturbed forecast F0, where F0 denotes the fake forecast fake000, etc. These cases are F3 (precipitation field shifted toward the SE), F6 [shifted as for F3 and with an additional multiplication by factor of 1.5: $R(F6) = 1.5 \times R(F3)$], and F7 [shifted as for F3 and with an additional subtraction of 1.27 mm: $R(F7) = \max(0, R(F3) - 1.27 \text{ mm})$]. The three modified cases (F3, F6, and F7) are shown in Fig. 2, where the red contour corresponds to the threshold contour $R^*$ for the identification of the precipitation objects (cf. section 2a).

![Synthetic precipitation fields, so-called geom-cases: (a) G0, (b) G2, (c) G3, and (d) G5. Lighter and darker shades of gray indicate light and intense precipitation.](image-url)
It is evident for F0, F3, and F6 that the identified objects in the interior of the domain are identical, because neither a shift of the precipitation field (F3) nor a linear scaling (F6) influence the identification of the objects. However, at the border of the domain, the displacement of the precipitation field will shift some objects out of or into the domain, and therefore the A and S components are not zero when comparing F0 versus F3 (see Table 2).

Particularly interesting is the comparison F0 versus F7: As expected when subtracting a constant value, F7 underestimates the total precipitation in the domain as revealed by the negative value of $A = -0.31$. However, the structure component also attains a similarly negative value ($S = -0.26$). This occurs because the precipitation objects become smaller when setting precipitation values below $R_0 = 1.27$ mm to zero (cf. Figs. 2a and 2c). This effect on $S$ when subtracting a constant value $R_0$ is particularly pronounced for strongly variable precipitation objects, for instance in a situation where large parts of the object receive weak stratiform precipitation and a small part receives intense convective precipitation. In such a situation (Fig. 3a), subtracting a constant value (which is assumed to be comparable to the stratiform precipitation amount) changes the size of the precipitation object considerably. In contrast, if the precipitation object’s precipitation distribution is relatively uniform (Fig. 3b), then the effects of subtracting a constant value are comparatively weak. In terms of verifying QPFs, this idealized example confirms that the structure component has the potential to distinguish between uniformly stratiform, mixed stratiform–convective, and purely convective precipitation fields. For instance, models that predict isolated convective cells in a situation where convective precipitation is embedded in widespread stratiform rain are characterized by negative values of $S$.

For a discussion and interpretation of the classical error measures for these fake cases (shown for completeness in Table 2), the reader is referred to AGBE.

5. Real QPFs from three models

The intercomparison exercise comprises forecasts from three versions of the WRF model, one with a horizontal resolution of 2 km (denoted here as 2CAPS) and two with a horizontal resolution of 4 km (4NCAR and 4NCEP). Each of these models was run without a parameterization of deep convection. For a discussion of further details of these model versions the reader is referred to Kain et al. (2008) and AGBE. Hourly accumulated QPFs from these models and radar-derived observational datasets have been available for 9 days in April–June 2005. These small samples do not allow for a systematic (statistical) evaluation of the models’ levels of SAL performance. Therefore, in this study the aim is not to identify the best model version, but rather to discuss a few selected QPFs, which are not necessarily representative for the three model versions, but which illustrate the information that can be obtained from SAL in comparison with classical error measures.

a. A selected example

First, the QPFs from all three models are considered for 1 June 2005. The main aim of this example is to emphasize the importance of prespecifying a meaningful domain when verifying QPFs with SAL. The precipitation distribution on this day was characterized by an

---

3 The QPFs correspond to the precipitation accumulated between forecast hours 23 and 24 of simulations started at 0000 UTC and the validation day will be used throughout this paper when referring to these QPFs.
elongated band (with distinct breaks) extending from North Dakota to northern Texas and weaker precipitation in Alabama and the neighboring states (Fig. 4a). First, SAL is applied to the QPFs in (almost) the entire domain with dimensions of about 2000 km x 1600 km (see large black rectangles in Fig. 4). In this large domain, the 95th percentile value corresponds to 6.81 mm h\(^{-1}\), and [cf. Eq. (1)] the threshold value for identifying precipitation objects is \( R^* = 0.454 \) mm h\(^{-1}\). The solid red contour in Fig. 4a corresponds to this threshold value and outlines the identified objects in the observational dataset (four large and several small objects). At first sight, all forecasts capture the precipitation band in the central part of the domain quite accurately. The 95th percentile values in the forecasts and the observations are comparable and therefore similar thresholds are used for identifying the objects. In the SE region it is apparent that the QPFs are characterized by many small objects (instead of one large one).

When calculating SAL for the large domain, the resulting values indicate that 4NCAR performs best on this day (Fig. 4c) with the smallest absolute values for the structure and amplitude components (see Table 3). The negative value of \( S = -0.22 \) points to the fact that the precipitation objects are slightly too small. The other QPFs have positive values for \( A \) and \( S \): 2CAPS (Fig. 4b) produces too much precipitation by mainly two large objects and 4NCEP (Fig. 4d) overestimates the precipitation amount with one very large object. For 4NCEP, the relatively large value of \( L = 0.29 \) (compared to the other models) is due to \( L_2 \), the second contribution to the location component, which measures the weighted averaged distance of the objects to the overall center of mass. Since this model essentially produces a single large precipitation object (instead of several roughly equally important objects), this distance is clearly too small and \( L_2 \) is therefore comparatively large.

The much too small structure of the QPF objects in the SE part of the domain is not well reflected by these values of \( S \) for the large domain, because objects with a small contribution to the total precipitation have only a weak influence on \( S \) [see Eqs. (8) and (9) in WPHF] and on this day the contributions from the objects along the frontal band clearly dominate. It is therefore more meaningful to apply SAL in more confined domains in order to separate the meteorologically differing precipitation systems in the central and northern parts of the considered domain, as well as along the south coast. If restricting the SAL analysis to a northern box (see red rectangles in the panels in Fig. 4), the values (not shown) are fairly similar as for the large domain, because (i) the overall precipitation maxima are located in the northern box and therefore the threshold for identifying objects is unchanged when going from the large to the northern box, and (ii) the most intense precipitation objects are all within the northern box. The main difference between the large and the northern box is that the \( L \) values become larger for the smaller domain, simply because the displacement errors in the calculation of \( L \) are scaled with the size parameter \( d \) of the domain (cf. section 2a).

However, totally different values for SAL are obtained when restricting the analysis to the small southern box (see Table 3). Figure 5 shows the identified objects in this southern domain. Since the maximum gridpoint value of precipitation in this domain is smaller than in the northern domain, the threshold value \( R^* \) is reduced when focusing on the southern domain, which renders the objects slightly larger in Fig. 5 compared to Fig. 4. But still it is obvious that the forecasts produce too many, too small, and too peaked precipitation objects, which

<table>
<thead>
<tr>
<th>Comparison</th>
<th>( S )</th>
<th>( A )</th>
<th>( L )</th>
<th>ME</th>
<th>RMSE</th>
<th>nRMSE</th>
<th>HSS</th>
<th>FBI</th>
</tr>
</thead>
<tbody>
<tr>
<td>F0 vs F3</td>
<td>0.03</td>
<td>-0.01</td>
<td>0.03</td>
<td>0.00</td>
<td>2.91</td>
<td>0.02</td>
<td>0.39</td>
<td>1.02</td>
</tr>
<tr>
<td>F0 vs F6</td>
<td>0.03</td>
<td>0.38</td>
<td>0.04</td>
<td>0.23</td>
<td>3.73</td>
<td>0.02</td>
<td>0.40</td>
<td>1.03</td>
</tr>
<tr>
<td>F0 vs F7</td>
<td>-0.26</td>
<td>-0.31</td>
<td>0.06</td>
<td>-0.13</td>
<td>2.82</td>
<td>0.02</td>
<td>0.27</td>
<td>0.51</td>
</tr>
</tbody>
</table>

**TABLE 2.** SAL values and classical error measures (see Table 1) for selected pairs of “fake” cases.

---

**Fig. 3.** Schematic cross sections showing the effects of subtracting a constant precipitation amount \( R_0 \) from (a) a peaked precipitation object and (b) a flat object.
are now well reflected by strongly negative $S$ values and positive values of $A$ (except for 4NCEP where $A$ is slightly negative). This strikingly different assessment of the QPF quality when considering the large domain or a particular subdomain indicates that the choice of the verification domain has a profound influence on the results of SAL (and on some of the classical error measures; see Table 3). If the domain is so large that it encompasses strongly differing meteorological systems, then the results might not be representative for the weaker precipitating system. It is therefore generally recommended that the domain size for the application of SAL should not exceed about $500 \times 500$ km$^2$. We will comment on this issue more generally in section 6.

b. Good and bad—According to classical error measures and SAL

In this section, the goal is to further elucidate the different assessments of QPF quality from classical measures and SAL. To this end forecasts have been selected for a closer analysis that have very small ME and RMSE values (Fig. 6), very large RMSE values (Fig. 7), and strongly contrasting quality in terms of the HSS (Fig. 8). The final examples (Fig. 9) present the QPFs that score best in terms of SAL.

According to the ME and RMSE, excellent QPFs are produced by the models 2CAPS and 4NCAR in the southern region on 3 June 2005 (see Table 3). For these QPFs, the ME is very close to zero and the RMSE is smaller than 1 mm. Also in terms of HSS and FBI, these QPFs are of at least moderate quality. However, Fig. 6 provides a different impression: although both forecasts capture the precipitation event in Chicago and its environment very well, they produce too small (and slightly too intense) precipitation objects in Tennessee and farther south. Accordingly, the $S$ values are strongly negative ($-0.79$ and $-0.92$, respectively) and indicate that the character of the precipitation events is not well
captured by the models. Also, the relatively large values of \( L \) (about 0.3 for both QPFs) point to the fact that the two forecasts are far from perfect. The \( A \) values are fairly small, in agreement with the small values of ME. One main reason for the small RMSE values is the overall low precipitation intensity of this event. This is also reflected by the medium values of the nRMSE (0.12 compared to \( \approx 0.07 \) for the QPFs presented for 1 June 2005). SAL also measures the relative errors of the QPFs and is therefore able to identify the forecast deficiencies of low-intensity events.

Figure 7 presents the two QPFs with the largest RMSE values of all QPFs. For both cases, which occurred on the same day in different domains, the RMSE amounts to \( \approx 4.3 \) mm. However, according to all other error measures, the quality of the two QPFs differs significantly (see Table 3). Figure 7 also indicates the strongly different character of the precipitation fields (and QPF errors) in the two cases. On 13 May 2005, an intense precipitation band was present in the northern domain, extending from Iowa to Kansas (Fig. 7a). The considered QPF (Fig. 7b) produced the shape and intensity of this band rather accurately but misplaced it by about 100 km to the east. In the second example with a very large RMSE the observations show weak precipitation in Mississippi (Fig. 7c), whereas the model produced much too intense precipitation, in particular, farther north in Tennessee (Fig. 7d). For practical purposes (i.e., for a weather forecaster or for hydrological applications), the first QPF (Fig. 7b) provides useful information (occurrence of an intense band of precipitation) whereas the second one (Fig. 7d) is in all aspects very far from reality. It is therefore not desirable that the quality of the two QPFs be assessed to be equal as is done by the RMSE. In the first example a large RMSE occurs because of the double-penalty problem, in the second case the occurrence is a result of the strong overestimation of the precipitation intensity and extent. The guidance provided by SAL appears more meaningful: in the first case, \( S \) and \( A \) are moderately positive (the band is slightly too large and too intense in the forecast) and \( L = 0.24 \) quantifies the shift of the band compared to the observations. For the second example, both \( S \) and \( A \) are very large, clearly indicating that the QPF produces too much rain in too large areas. The poor quality of this QPF is also indicated by the large values of ME, nRMSE, and FBI and the almost zero HSS.

As a next example, Fig. 8 illustrates two QPFs with strongly contrasting quality in terms of the HSS. Again, they occurred on 13 May 2005; one each in the northern and southern domains. For the rainband in the northern domain (Figs. 8a and 8b), the model 4NCAR attained the largest HSS of all QPFs in the dataset because of a large overlap of the observed and predicted precipitation structures. The SAL values indicate a moderate level of quality by the forecast, where \( L = 0.24 \) is mainly due to missing smaller objects in the north and southwest of the domain. The positive value of \( A \) indicates a moderate overestimation of the precipitation amount, whereas the relatively large value of \( S = 0.74 \) might be surprising at first glance. The reason for this value is the small but intense object in northern Texas (see Fig. 8a) that was missed by the forecast. Its maximum value is larger than 70 mm h\(^{-1}\) (whereas the maximum values of the observed and predicted precipitation bands farther to the north amount to 40–50 mm h\(^{-1}\)). Therefore, the northern Texas precipitation object is strongly peaked leading to a small “scaled volume” [cf. Eq. (4)] and hence (note that this peaked object is missed by the forecast) to a relatively large positive value of the structure component \( S \).
In contrast, the HSS of the QPF of the same model in the southern domain (Figs. 8c and 8d) is almost zero, indicating a nonskillful forecast. In addition, the nRMSE shows a fairly large value compared to most other cases. Unlike for the erroneous QPF shown in Fig. 7d (which had an equally poor HSS), visual inspection indicates a useful forecast quality, in the sense that the QPF captured the character of the event with several moderate showers in Mississippi and Tennessee (cf. Figs. 8c and 8d). The fact that the locations of the many small objects in the observations and forecast do not coincide is responsible for the very low HSS. The information obtained from SAL for this case is again more meaningful: $S$ is slightly negative, indicating in this case that the objects are too peaked (too large maximum values in the small objects). Accordingly, $A$ is fairly large, pointing to an overestimation of the total rainfall. Also here, as discussed for the examples shown in Fig. 5, the values of the ME, RMSE, and FBI are small, mainly because of the low precipitation amount of the event.

Finally, the two best forecasts in terms of SAL are presented in Fig. 9. They have been selected as the forecast in either the northern or southern domain with the smallest absolute values of all SAL components, that is, as the QPF for which $\max(|S|, |A|, L)$ is smallest.4 Within the set of QPFs used in this study, the best QPFs have values of $\max(|S|, |A|, L)$ equal to 0.19 (2CAPS in the northern domain on 22 April; Figs. 9a and 9b) and 0.22 (4NCAR in the northern domain on 1 June; Figs. 9c and 9d), respectively (see again Table 3). These values are not particularly small—Hofmann et al. (2009) present a few examples of excellent QPFs over Germany with values of $\max(|S|, |A|, L)$ less than 0.1—indicating

---

4 This is a reasonable choice for defining a scalar metric from the three SAL components, which attributes high quality to QPFs that are characterized by small values of all SAL components. In contrast, an alternative scalar measure like $(S^2 + A^2 + L^2)^{0.5}$ would allow compensating (e.g., a larger value of $L$ by small values of $S$ and $A$).
that the best QPFs considered here also have some noticeable deficits. Figure 9b shows that the forecast captured well the precipitation band extending along the northern and eastern borders of Nebraska (cf. with Fig. 9a), although with a certain shift to the west. For the second-best QPF (according to SAL), Figs. 9c and 9d illustrate the relatively high quality of the QPF capturing the patchy precipitation band extending from northern Texas to South Dakota, however with significant displacements of the precipitation objects on smaller scales. It is interesting to see that for both examples the classical error measures provide a rather mixed picture of the quality of these QPFs (see Table 3): the ME, nRMSE, and FBI values are close to optimal for both cases, the
RMSE values are intermediate, and the HSS values are excellent for the first (0.41) but rather poor for the second (0.17) case. The reason might be that the smaller-scale displacements of the precipitation objects in the second case lead to a limited overlapping of the predicted and observed objects and hence to a relatively small HSS. These final examples indicate that a high quality QPF according to SAL does not necessarily score very well for all classical error measures.

FIG. 8. Precipitation fields on 13 May 2005. Observations in the (a) northern and (c) southern domains, and forecasts from 4NCEP in the (b) northern and (d) southern domains.

FIG. 9. Precipitation fields in the northern domain on 22 Apr 2005 for (a) observations and (b) forecast from 2CAPS, and on 1 Jun 2005 for (c) observations and (d) forecast from 4NCAR.
6. Conclusions

The QPF cases over the United States, which have been selected as sample cases for the verification method intercomparison project, provide useful additional insights into the behavior of the SAL quality measure, in particular when comparing with classical error measures. The most important findings can be summarized as follows:

- Investigation of the geometric cases supports the findings of WPHF—that SAL captures errors in the location and size of well-defined single precipitation objects, but that SAL is not sensitive to errors in the orientation of precipitation objects.
- Similar conclusions can be drawn when analyzing the fake cases. In addition, it was found that reducing the precipitation by subtracting a constant value leads to smaller and more peaked precipitation objects, which are well captured by a significantly negative value of the structure component $S$. This indicates that SAL (in particular $S$) has the potential to distinguish between forecasts characterized by isolated intense precipitation objects and forecasts characterized by intense precipitation embedded within a larger-scale low-intensity precipitation area.
- The WRF model QPFs over the United States occasionally revealed contrasting behavior for the SAL compared to classical error measures. As expected, absolute error measures like ME and RMSE are typically good for weak events; in contrast, the relative measures used by SAL are often better for intense events.
- High quality QPFs, according to SAL, are not necessarily associated with good values of the classical error measures; in particular, the HSS can indicate a relatively poor performance for cases that are rated well by SAL.
- Overall, the detailed investigation of the idealized and real intercomparison project QPFs indicates that SAL provides useful and meaningful guidance about the quality and specific deficiencies of a particular forecast.

In addition, it was found that the application of SAL in the full domain with an extension of about 2000 km $\times$ 1600 km (cf. section 5a) might lead to results that are difficult to interpret, in particular if the domain contains several meteorologically distinct precipitation systems (e.g., a frontal rainband and a region of air mass convection). In such a situation QPF errors might differ for the precipitation systems and the resulting SAL values might correspond to an average that is not necessarily meaningful. It should be noted, however, that this is not a specific issue of the SAL technique since other error measures can suffer from the same cancellation and/or averaging effects if the verification domain becomes very large compared to the scale of the precipitation system. For instance, an underestimation of convective activity in one part of the domain can be compensated by an overestimation of stratiform rain in another part of the domain leading to an almost perfect ME value. For routine verification, it is very difficult to automatically determine a meaningful size of the verification domain and anyone interpreting these verification results should be aware of this potential caveat in case of large verification domains.

Finally, it should be noted that here (in contrast, e.g., to the study by WPHF) SAL has been applied to hourly accumulated QPFs. One of the outstanding challenges of today’s QPF environment is verification for such short accumulation times, during which precipitation fields are often characterized by very complex structures (e.g., enhanced patchiness and smaller-scale features compared to daily accumulated fields). It will be important to apply SAL and other recently developed QPF quality measures to larger sets of QPFs with short accumulation times.

Acknowledgments. MZ acknowledges funding from the German Research Foundation (DFG) priority program on Quantitative Precipitation Forecasts (SPP 1167).

REFERENCES


