A Technique for the Verification of Precipitation Forecasts and Its Application to a Problem of Predictability

FAN HAN AND ISTVAN SZUNYOGH

Texas A&M University, College Station, Texas

ABSTRACT

A new morphing-based technique is proposed for the verification of precipitation forecasts for which the location error can be described by a spatial shift. An adaptation of the structural similarity index measure (SSIM) of image processing to the precipitation forecast verification problem, called the amplitude and structural similarity index (ASSIM), is also introduced. ASSIM is used to measure both the convergence of the new morphing algorithm, which is an iterative scheme, and the amplitude and structure component of the forecast error. The behavior of the proposed technique, which could also be applied to other forecast parameters with sharp gradients (e.g., potential vorticity), is illustrated with idealized and realistic examples. One of these examples examines the predictability of the location of precipitation events associated with winter storms. It is found that the functional dependence of the average magnitude of the location error on the forecast lead time is qualitatively similar to that of the root-mean-square error of the fields of the conventional atmospheric state variables (e.g., geopotential height). Quantitatively, the average magnitude of the estimated location error is about 40 km at initial time, 110 km at day 1, 250 km at day 3, and 750 km at week 1, and it eventually saturates at about week 2.

1. Introduction

In a pair of papers, Keil and Craig (2007, 2009; hereafter KC07 and KC09, respectively) introduced a morphing-based, nonparametric optical flow technique (Marzban et al. 2009) for the verification of precipitation forecasts. Their technique was most recently used by Geiß (2015) to examine the forecast cases of the Mesoscale Verification Intercomparison over Complex Terrain (MesoVICT) research project (Dorninger et al. 2013). In an earlier paper (Han and Szunyogh 2016; hereafter HS16), we proposed some algorithmic changes to the morphing technique of KC07 and KC09 and also defined the morphing-based estimate of the components of the forecast error differently. The primary motivation for the changes we introduced was to improve the performance of the technique in situations where the precipitation in the verification region is produced by a single weather event, such as the passage of the frontal system of an extratropical cyclone, a tropical cyclone, or a form of organized convection. In such situations, the error in the prediction of the location of the precipitation system is well defined and can be described by a spatial shift.

The purpose of the present paper is to introduce further algorithmic changes to the morphing technique in order to improve its efficiency in matching the forecast and verifying precipitation features. The paper also introduces a measure of the similarity between a pair of precipitation fields, which we call the amplitude and structural similarity index measure (ASSIM). ASSIM is used both to define the convergence criterion for the revised morphing technique, which is now an iterative algorithm, and to measure the amplitude and structure error of the forecasts. The formal definition of ASSIM derives from that of the universal quality index (UQI) of Wang and Bovik (2002), sometimes called the Wang–Bovik index. We call the measure ASSIM, in part to acknowledge its formal similarity to the best-known variant of the Wang–Bovik index, the structural similarity index measure (SSIM; Wang et al. 2004; Wang and Bovik 2009), and in part to emphasize that it accounts for both the amplitude and the structure error of precipitation forecasts.

The structure of the paper is as follows. Section 2 summarizes the morphing-based technique of HS16, describes the proposed changes to the technique, and introduces ASSIM. Section 3 illustrates the proposed verification approach with the help of idealized and realistic test cases. Section 4 is an application of the proposed verification scheme to the assessment of the predictability of precipitation events

---

Corresponding author: Fan Han, hanfan5598@gmail.com

associated with winter storms. An extension of our verification technique to ensemble forecasts of the same winter storms is presented in Han and Szunyogh (2018). Section 5 offers our conclusions.

2. The proposed verification technique

The proposed verification technique treats all precipitation in the verification domain as part of a single precipitation system. We introduce the technique in three steps. First, we briefly describe the basic pyramid matching algorithm of KC07 and KC09 (section 2a). Then, we summarize the changes that HS16 made in the implementation of the algorithm and the interpretation of the morphing results (section 2b). Finally, we introduce a new high-level algorithm, which uses the HS16 implementation of the pyramid matching algorithm as a lower-level algorithmic component (section 2c). We consider this high-level algorithm and the introduction of ASSIM (section 2d) the most important novel aspects of the present study.

a. The pyramid matching algorithm

Assume that the precipitation fields are represented by a collection of $d \times d$ elementary pixels: each pixel represents the areal mean of the total precipitation for the pixel for a specific time interval. Given a forecast precipitation field P^f and a corresponding verifying analysis field P^a, the goal is to morph the image of P^f into an approximate image $\hat{P}^f$ of P^a. The morphing is carried out by computing a morphing vector for each (elementary) pixel of P^f that moves the pixel to its new location in the morphed image. If no other pixel moves into the original location of the pixel, a new zero pixel is placed at its original location. The vector field defined by the morphing vectors is an optical flow.

The number of pyramid levels F, the free parameter of the algorithm, is called the subsampling parameter. At level k ($k = F, F - 1, \ldots, 0$) of the algorithm, the forecast and the verifying precipitation fields are coarse grained by averaging $2^k \times 2^k$ (elementary) pixels of both fields. Then, each pixel of the coarse-grained forecast image is shifted by 0, 1, and -1 position in both the zonal (x) and meridional (y) directions. The dx zonal and dy meridional components of the morphing vector $d\mathbf{X}$ are chosen to be the pair of values of 0, 1, and -1 that minimizes the absolute value of the difference between the coarse-grained forecast pixel and the coarse-grained verifying analysis pixel. (The same $d\mathbf{X}$ is assigned to all $2^k \times 2^k$ elementary pixels that make up a coarse-grained pixel, but $d\mathbf{X}$ can be different for the different coarse-grained pixels.) The procedure is repeated until its completion for the last level (level 0), at

which the coarse-grained pixels are identical to the elementary pixels ($2^0 \times 2^0 = 1$). The maximum total distance that an elementary pixel of the forecast field can move as the result of morphing in either direction is $2^F \times d$.

b. The morphing technique of HS16

There were a number of important differences in the technical details of the implementation of the pyramid matching algorithm by KC07 and KC09 versus HS16. In particular, unlike KC07 and KC09, HS16

1) imposed a constraint at the top level ($k = F$) of the pyramid to prevent moving pixels of P^f in which the precipitation was not likely to be related to a precipitation feature of P^a (section 2c of HS16);
2) merged elementary pixels that arrived at the same location in order to preserve the total precipitation; and
3) did not require the specification of F, as it chose the optimal value of F based on P^f and P^a (section 3b of HS16).

The technique of the present paper (section 3c) retains properties 1 and 2 of HS16. As for property 3, it should be noted that while the technique of HS16 did not require the specification of F, it required the specification of the maximum allowable value of F, F_{max}. The main motivation to cap the value of F was twofold. First, it limited the distance $L_{max} = 2^{F_{max}} \times d$ in both the zonal and meridional directions, within which precipitation features could be considered matching. Second, it reduced the likelihood of the presence of multiple precipitation systems within the search distance. Limiting the value of F, however, had an unintended consequence: it led to an underestimation of the location error in situations where only part of the forecast feature was within distance L_{max} from the verifying feature. In section 2c, we will propose a number of modifications to the morphing technique of HS16 to reduce the effects of this undesirable behavior of the technique. The proposed technique is also computationally more efficient than that of HS16, which is an important advantage when a large number of cases has to be processed.

Besides changes to the implementation of the pyramid matching algorithm, HS16 also used the output of the algorithm (the morphing vectors) differently than KC07 and KC09 to define the components of the forecast error. While KC07 and KC09 defined the location error by the mean of the absolute value of the morphing vectors, HS16 defined the location error by the mean of the morphing vectors. That is, in practical terms, while HS16 measured the location error by only the translational component of the optical flow, KC07 and KC09 considered the nontranslational

components also part of the location error. HS16 argued that some of those components (e.g., the divergent component) indicated an error in the structure rather than the location of the precipitation feature, while some other components were pure artifacts of the pyramid matching algorithm.

Another important difference between the error measures of HS16 and KC07 and KC09 was that the measure of HS16 had three components, while the measures of KC07 and KC09 had only two components. In particular, HS16 defined a structure and amplitude error component in addition to the location error component, similar to Wernli et al. (2008), which was computed after a correction for the location error by shifting P^f. In contrast, KC07 and KC09 measured the nonlocation errors by a single residual error term.

In the measure proposed by the present paper, the location error is defined as in HS16, except that the morphing vectors are computed differently, while the structure and amplitude error component is defined differently.

c. The modified morphing technique

The steps of the proposed algorithm are the following:

1) Filtering. Prepare a filtered image P^f_{filtered} of P^f by replacing each nonzero pixel of P^f that is at least $L_{\max}$ pixels apart from the nearest nonzero pixel of P^a in either direction by a zero pixel. This step is applied to the elementary pixels.
2) Morphing. Use the pyramid matching algorithm to morph P^f_{filtered} into an approximate image $\hat{P}^f$ of P^a to obtain the field of morphing vectors $d\mathbf{X}_{i,j}$, where $i = 1, 2, \ldots, M$ and $j = 1, 2, \ldots, N$ identify the location of a pixel of P^f_{filtered} in the x and y direction, respectively.
3) Shifting. Compute the mean,

$$d\mathbf{X}_{\text{mean}} = \frac{1}{n} \sum_{i=1}^{M} \sum_{j=1}^{N} d\mathbf{X}_{i,j}, \qquad (1)$$

of the morphing vectors, where n is the number of nonzero pixels of P^f_{filtered}. Round the dx_{mean} and dy_{mean} components of $d\mathbf{X}_{\text{mean}}$ to the nearest integers, and shift all nonzero pixels of P^f by $d\mathbf{X}_{\text{mean}}$. [Note that $d\mathbf{X}_{\text{mean}}$ is a pure translational (constant) vector field that has no divergent, rotational, or deformation component.]
4) Iteration. Replace P^f by the shifted field and repeat steps 1–3 until the shifted field matches P^a as closely as possible. A formal criterion for "matching P^a as closely as possible" will be provided in section 2d.

The key modification of the morphing process is the repeated (iterative) application of the pyramid matching algorithm and the shift vector $d\mathbf{X}_{\text{mean}}$. (In HS16, the forecast feature was shifted by $d\mathbf{X}_{\text{mean}}$ only once, after

the completion of the morphing process, as it was part of the computation of the measure of the structure error rather than the morphing technique.) The final shifted forecast field $d\mathbf{X}_{\text{final}}$ can be equivalently computed by applying the sum of $d\mathbf{X}_{\text{mean}}$ from all iterations to the pixels of the original forecast field P^f. Because $d\mathbf{X}_{\text{final}}$ is a sum of constant vector fields, it is itself a constant vector field that shifts all elementary pixels of P^f to the same distance and in the same direction. While this algorithm still requires prescribing $F_{\max}$, it can move pixels to distances significantly longer than $L_{\max}$. This feature is highly advantageous when the problem at hand is to estimate the large errors of medium- and long-range forecasts.

d. ASSIM

As mentioned earlier, ASSIM is an adaptation of the image quality measure UQI (Wang and Bovik 2002) to the precipitation forecast verification problem. We use ASSIM to

- provide a formal definition of "matching P^a as closely as possible given P^f" in step 4 of the morphing technique (section 2c),
- measure the amplitude and structure error, and
- match forecast and analysis features.

For a pair of precipitation fields P^1 and P^2, UQI is

$$U(P^1, P^2) = [U_1(P^1, P^2)]^\alpha \times [U_2(P^1, P^2)]^\beta \times [U_3(P^1, P^2)]^\gamma, \qquad (2)$$

where α, β, and γ ($\alpha + \beta + \gamma = 1$) are user-defined parameters. The larger the one of these parameters, the larger the relative weight of the associated component in the index. The functions $U_1(P^1, P^2)$, $U_2(P^1, P^2)$, and $U_3(P^1, P^2)$ are defined by

$$U_1(P^1, P^2) = \frac{2\mu_1\mu_2}{\mu_1^2 + \mu_2^2}, \qquad (3)$$

$$U_2(P^1, P^2) = \frac{2\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2}, \qquad (4)$$

$$U_3(P^1, P^2) = \frac{\sigma_{12}}{\sigma_1\sigma_2}, \qquad (5)$$

where $U_1(P^1, P^2)$ is a measure of the similarity between the amplitudes of P^1 and P^2, $U_2(P^1, P^2)$ is a measure of the similarity between the variances (spatial variability) of P^1 and P^2, and $U_3(P^1, P^2)$ is a measure of the pointwise linear correspondence between P^1 and P^2.

The term $[U_1(P^1, P^2)]^\alpha$ is a measure of the amplitude error, and the product $[U_2(P^1, P^2)]^\beta \times [U_3(P^1, P^2)]^\gamma$ is a measure of the structure error; $U(P^1, P^2)$ always takes a value in the closed interval $[1, -1]$, and when P^1 and P^2 are identical, $U(P^1, P^2) = 1$. UQI measures the same

three aspects of the difference between P^1 and P^2 as the root-mean-square (RMS) error, but it better reflects human visual perception. UQI and SSIM were originally introduced as objective measures of perceptual image quality. By adapting a UQI-based measure of the similarity, rather than using the RMS difference, we hope to achieve a better agreement with the assessment of human experts about the similarity of the two fields.

We define ASSIM by

$$D(P^1, P^2) = [D_{\mathrm{amp}}(P^1, P^2)]^\alpha \times [D_{\mathrm{var}}(P^1, P^2)]^\beta$$
$$\times [D_{\mathrm{cor}}(P^1, P^2)]^\gamma, \qquad (6)$$

where the three components are

$$D_{\mathrm{amp}}(P^1, P^2) = \begin{cases} 1, & \text{if } \mu_1 + \mu_2 < \varepsilon, \\ U_1(P^1, P^2), & \text{otherwise;} \end{cases} \qquad (7)$$

$$D_{\mathrm{var}}(P^1, P^2) = \begin{cases} 1, & \text{if } \sigma_1^2 + \sigma_2^2 < \varepsilon^2, \\ U_2(P^1, P^2), & \text{otherwise;} \end{cases} \qquad (8)$$

$$D_{\mathrm{cor}}(P^1, P^2) = \begin{cases} 1, & \text{if } \sigma_1 \text{ or } \sigma_2 < \varepsilon, \\ U_3(P^1, P^2), & \text{otherwise, if } \sigma_{12} \geq 0, \\ 0, & \text{otherwise, if } \sigma_{12} < 0. \end{cases} \qquad (9)$$

The parameter ε is a prescribed small threshold value, below which the estimates of the statistical parameters, which are based on computer-generated precipitation fields, should be considered zero. When ASSIM is used as a measure of forecast error, P^1 is a forecast, and P^2 is the verifying analysis.

There are two differences between the definitions of ASSIM and UQI. First, they treat the special cases $\mu_1 = \mu_2 = 0$ and $\sigma_1 = \sigma_2 = 0$ differently: unlike UQI, ASSIM has no singularity for these special cases. We note that SSIM (Wang et al. 2004) was also introduced to remove the singularity associated with the same special cases by adding a small additive constant to both the numerator and the denominator of the right-hand side of Eqs. (3)–(5). We choose a different approach to remove the singularity because both $\mu_1 = \mu_2 = 0$ and $\sigma_1 = \sigma_2 = 0$ have a clear interpretation in the context of precipitation fields: the former is the case of no precipitation in either field, while the latter is the case of spatially uniform precipitation in both fields. In the context of forecast verification, the former is a perfect prediction of the amplitude, while the latter is a perfect forecast of the spatial variability.

The second difference between ASSIM and UQI is that in the definition of $D_{\mathrm{cor}}(P^1, P^2)$, the negative correlation values σ_{12} are replaced by zeros. One motivation for this modification is the convention of meteorology to define skill scores, such that they take a value in the closed interval [1, 0], with one indicating a perfect

forecast and zero a forecast with no skill. Another, more important, motivation is to ensure that a lower value of the measure indicates a poorer forecast. If negative values of $D_{\mathrm{cor}}(P^1, P^2)$ were allowed, this requirement would not be satisfied, because for a negative value of $D_{\mathrm{cor}}(P^1, P^2)$, a more accurate forecast of the amplitude or the variance (a larger value of $D_{\mathrm{amp}}(P^1, P^2)$ or $D_{\mathrm{var}}(P^1, P^2)$) would result in a smaller value (a negative number with a larger absolute value) of $D(P^1, P^2)$.

Similar to UQI (Wang and Bovik 2002) and SSIM (Wang et al. 2004), ASSIM satisfies the following conditions:

- Symmetry: $D(P^1, P^2) = D(P^2, P^1)$.
- Boundedness: $0 \leq D(P^1, P^2) \leq 1$.
- Unique maximum: $D(P^1, P^2) = 1$, if and only if $P^1_{i,j} = P^2_{i,j}$ for $i = 1, \ldots, M, j = 1, \ldots, N$.

When ASSIM is used for forecast verification, a location error reduces ASSIM because it reduces $D_{\mathrm{cor}}(P^f, P^a)$. Thus, a morphed forecast field $\hat{P}^f$ matches P^a as closely as possible, given P^f, when $D(\hat{P}^f, P^a)$ takes its maximum, given P^f and P^a. In practical terms, the iterations of the morphing technique can be stopped once an iteration fails to increase ASSIM by more than a prescribed small value ν.

Finally, we note that in the atmospheric sciences, UQI was considered before as a similarity index by Venugopal et al. (2005) and Mo et al. (2014). In particular, Venugopal et al. (2005) combined the components $U_1(P^1, P^2)$ and $U_2(P^1, P^2)$ with a Hausdorff measure of the distance between the locations of precipitation features to define a new measure for the verification of precipitation forecasts. They did not use $U_3(P^1, P^2)$ because their goal, similar to ours, was to separate the amplitude and structure error[1] from the location error, and as they observed, $U_3(P^1, P^2)$ is affected by both types of errors. They noted that $U_3(P^1, P^2)$ could be used to measure the structure error if there were an objective technique to find the shift that maximizes the similarity between the two features to be compared. In essence, our morphing technique is such an objective technique, allowing for the inclusion of $U_3(P^1, P^2)$ in the measure of the structure error.

3. Illustration of the verification technique

This section illustrates the behavior of the proposed verification technique by applying it to an idealized case of HS16, idealized and realistic cases of the Spatial Verification Methods Intercomparison Project (ICP) (Ahijevych et al. 2009), and realistic cases of the

---

[1] They referred to the amplitude and structure error collectively as amplitude error.
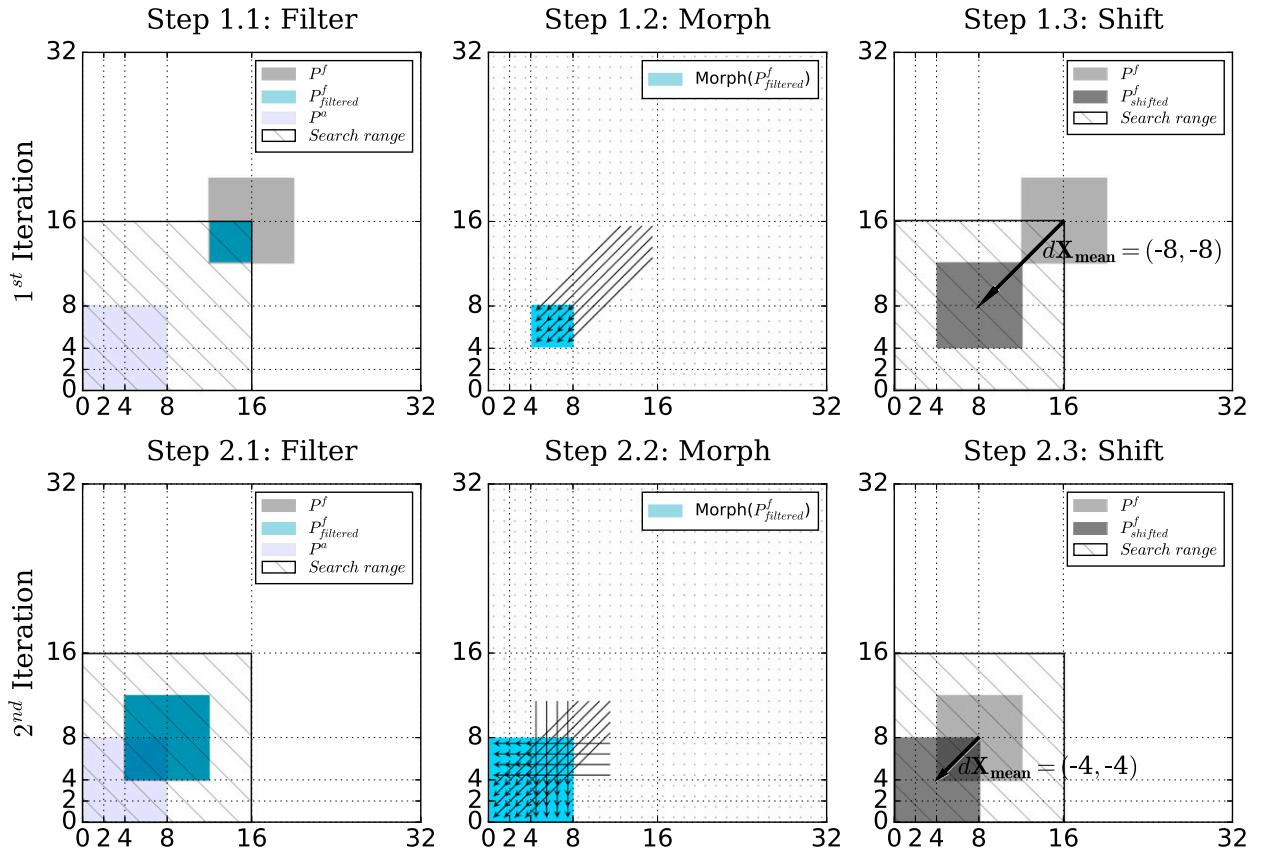
FIG. 1. Illustration of the steps of the proposed morphing technique for an idealized case in which the forecast has only location error (see section 3a for details). In this example, the forecast feature is misplaced by 12 pixels in both directions.

MesoVICT project. In our calculations, we assume that the three properties of the amplitude and structure errors are equally important for the hypothetical forecast user by choosing the values of the parameters in ASSIM to be $\alpha = \beta = \gamma = 1/3$. The values of ν and ε are, respectively, chosen to be 0.01 and 0.1 in all examples.

a. Idealized case of HS16

This case includes a series of idealized forecasts with increasing location error. The precipitation feature is an 8×8 pixel square, located in the bottom-left corner of the verification region (Fig. 1). The series of simulated forecasts is generated by gradually shifting the precipitation feature of P^f (one pixel at a time) in both the x and y directions. The panels of Fig. 1 illustrate the effect of each step of the first two iterations of the morphing technique for the forecast, in which the precipitation feature is misplaced by 12 pixels in both directions. In this calculation, we make the choice $F_{max} = 3$, and the iteration is stopped once ASSIM becomes 1. The first filtering step (top-left panel) filters the 3/4 part of the

forecast precipitation feature that is outside of the search region, that is, beyond a $L_{max} = 2^{F_{max}} = 8$ pixel distance from the border of the analysis feature in either the x or y direction. The first application of the pyramid matching algorithm (top middle) moves the remaining 1/4 of the forecast precipitation feature toward the verifying precipitation feature, while the first shift step (top right) moves the entire forecast precipitation feature into the search region. The steps of the second iteration (bottom panels) result in a perfect match of the two precipitation features.

Figure 2 illustrates the change in the behavior of the morphing technique due to the proposed modifications. It shows that when the forecast feature is misplaced by $2^{F_{max}} = 8$ or fewer pixels, both techniques capture the location error correctly. But, while the modified technique provides a correct estimate of the location error up to a displacement of 15 pixels, the original technique underestimates the location error at a displacement of nine or more pixels because it cannot move those parts of the precipitation feature that are displaced by more than $2^{F_{max}}$ pixels (Fig. 3).
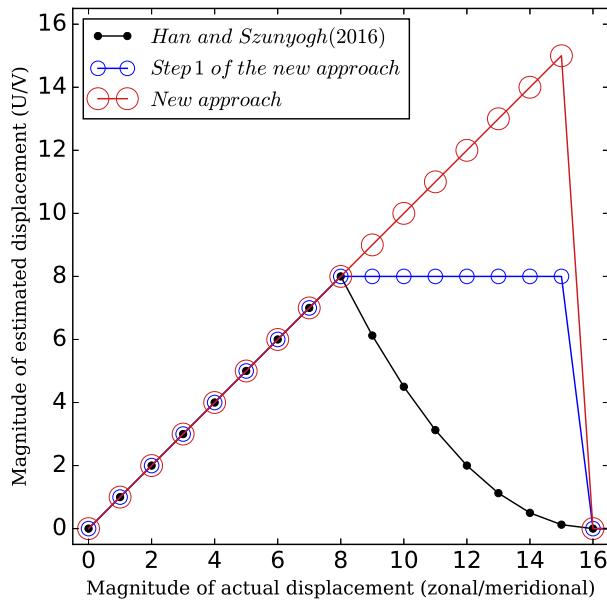
FIG. 2. The estimates of the displacement errors for the series of simulated forecasts described in section 3a that were obtained by the technique of HS16 (black), the first iteration of the proposed modified technique (blue), and the proposed modified technique (red).

b. ICP cases

1) IDEALIZED CASES

The proposed verification scheme is applied to the five idealized ICP examples. In these examples, whose graphical illustration can be found in Ahijevych et al. (2009), both the forecast and analysis features have elliptical shapes. The forecast errors are errors in the location, area, and/or aspect ratio of the ellipsoids. Table 1 summarizes the estimates of $d\mathbf{X}_{\text{final}}$, ASSIM, and the three components of ASSIM for these examples. The results show that our technique can distinguish among the three general types of error present in all cases.

1) When the forecasts only have location error (cases 1 and 2), the verification technique does a perfect job: the estimate of the location error $d\mathbf{X}_{\text{final}}$ equals the actual displacement, and both ASSIM and its components correctly indicate that there is no amplitude or structure error.

2) When both location and structure error are present (cases 3, 4, and 5), the estimates of the location error are slightly less accurate. In particular, while the technique correctly detects that there is no location error in the y direction for cases 3 and 4, it estimates the location errors in the x direction with a relative error of 3.2%.

3) When there is no error in the amplitude and the only source of structure error is an error in the aspect ratio (case 4), the components of ASSIM indicate correctly that there is no error in the amplitude or variance. In addition, the error in the aspect ratio is captured by the component D_{cor}, as it leads to less than perfect pointwise correlation. Unlike the verification techniques of Lack et al. (2010) and Lakshmanan and Kain (2010), it cannot detect specifically that the structure error is an error in the orientation of the forecast feature.

Table 2 ranks the forecasts of the five idealized cases based on different verification metrics. The first six rows show the rankings with respect to six modern verification metrics that can detect location error (Ahijevych et al. 2009; Davis et al. 2009; Lack et al. 2010; Keil and Craig 2009; Gilleland 2011; Lakshmanan and Kain 2010; Zhu et al. 2011), while the last three rows show results for our verification technique. For the preparation of this table, ASSIM was computed in two different ways: D^0 is ASSIM before the correction for the location error, and D is ASSIM after correction for the location error. We use the results of Table 2 to address three questions posed by Ahijevych et al. (2009).

(i) Does case 1 score better than case 2, and is the error correctly attributed to displacement?

Except for the traditional metrics and the two metrics based on ASSIM, the metrics of Table 2 indicate the "forecast" of case 1 is better than that of case 2. Except for the displacement and amplitude score (DAS; Keil and Craig 2009), for which the location error is not correctly detected in case 2, the better scores and ranks for case 1 are due to the correct detection of the presence of a smaller location error in case 1. This indicates that when location error is the only source of error present, most modern techniques (e.g., neighborhood, object-based, and morphing-based methods) are able to detect the location error and discriminate among cases with different magnitudes of the location error. The results for ASSIM support our argument that ASSIM should be computed after the correction of the location error and only be used as a component of a multivariate measure that has another component to measure the location error.

(ii) Is the method sensitive to the increasing frequency bias in case 3 and case 5?

Here, the term "increasing frequency bias" refers to the property of cases 3 and 5 that precipitation is forecast at more locations than observed. To be precise, the forecast features of cases 3 and 5 are stretched in the x direction, such that they lead to a frequency bias, which is larger in case 5 than case 3. There is also a 125-points

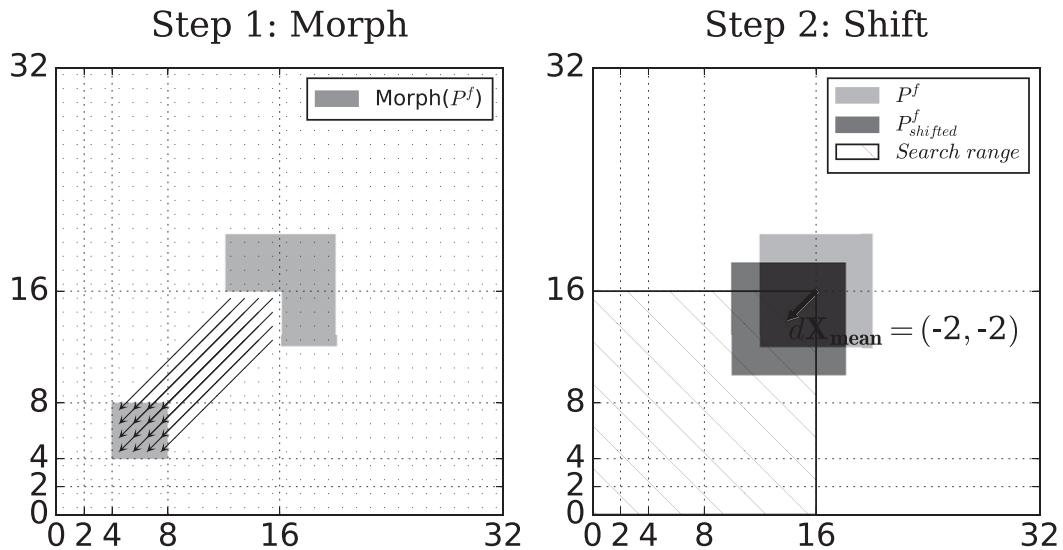## Step 1: Morph   ## Step 2: Shift

FIG. 3. Illustration of the steps of the morphing technique of HS16 for the idealized example of Fig. 1.

location error in both cases. A major source of difficulty for the verification techniques in this comparison is that the forecast and verifying feature partially overlap in case 5 because of the larger frequency bias, which somewhat masks the location error. The metrics Method for Object-based Diagnostic Evaluation (MODE; Davis et al. 2009), DAS, $|d\mathbf{X}_{\text{final}}|$, and ASSIM indicate, incorrectly, that the forecast of case 5 is better than that of case 3. In the case of $|d\mathbf{X}_{\text{final}}|$, it indicates that the forecast of case 5 is better than that of case 3 due to a small error in the estimate of the location error for case 3. This result shows that the location error should only be used as a component of a multivariate measure when the goal is to measure overall forecast quality. The results also support our earlier argument that while ASSIM should not be used as a measure of forecast quality without a prior correction for the location error, it can correctly discriminate among cases of different frequency bias after a correction for the location error.

A more complex comparison is between cases 2 and 5: the forecast feature of case 2 is identical to the verification feature, while the forecast feature of case 5 has a large

frequency bias, but case 2 has a larger location error. Five out of the nine metrics (MODE, Baddeley's Δ, metr$_v$, $|d\mathbf{X}_{\text{final}}|$, and ASSIM with no correction of the location error) indicate that the forecast of case 5 is better than the forecast of case 2. This result demonstrates that while it is more convenient to measure forecast performance by a scalar than a multivariate score, a scalar score often provides incomplete information about the errors. For instance, some users of the forecast information would find the larger location error of case 2 the more serious problem, while others would say that it is the larger frequency bias of case 5. But, if $|d\mathbf{X}_{\text{final}}|$ and ASSIM (or its components) are used together in a multivariate measure as we propose, the location error component correctly indicates that the location error is larger in case 2, while ASSIM correctly indicates that the frequency bias is larger in case 5.

(iii) Can the method diagnose the aspect ratio error in case 4?

As discussed earlier, ASSIM captures the aspect ratio error as a general structure error, but cannot specify its type.

TABLE 1. Summary of the results for the idealized ICP cases.

Case	Description of error	$d\mathbf{X}_{\text{final}}$ (pixel, pixel)	D_{amp}	D_{var}	D_{cor}	D
geom001	50-pixel displacement in x direction	$(-50, 0)$	1	1	1	1
geom002	200-pixel displacement in x direction	$(-200, 0)$	1	1	1	1
geom003	125-pixel displacement in x direction, area error	$(-129, 0)$	0.468	0.812	0.491	0.571
geom004	125-pixel displacement in x direction, wrong aspect ratio	$(-121, 0)$	1	1	0.31	0.677
geom005	125-pixel displacement in x direction, large area error but overlapping	$(-125, 0)$	0.245	0.661	0.301	0.365

TABLE 2. Rankings of the forecasts of the idealized ICP cases with respect to different verification metrics. Asterisks indicate identical ranks for multiple cases in the same row (for the same metrics).

Metric	Reference	geom001	geom002	geom003	geom004	geom005
MODE	Davis et al. (2009)	1	5	3	4	2
Procrustes	Lack et al. (2010)	1	2	3	4	5
DAS	Keil and Craig (2009)	1	2	5	4	3
Baddeley's Δ	Gilleland (2011)	1	5	3	2	4
e	Lakshmanan and Kain (2010)	1	2	4	3	5
$metr_v$	Zhu et al. (2011)	1	5	3	2	4
$|d\mathbf{X}_{final}|$		1	5	4	2	3
D^0		2*	2*	2*	2*	1
D		1*	1*	4	3	5

2) REALISTIC CASES

The proposed verification scheme is applied to the 27 ICP cases. A useful feature of the ICP dataset is that subjective scores are available for all 27 cases (Ahijevych et al. 2009). These subjective scores are the averages of scores assigned by each member of a panel of 26 scientists. Because the panel included meteorologists, statisticians, and software engineers, the scores should be considered more a reflection of human perception about the similarity between images than a consensus of trained experts on forecast quality.

Before turning our attention to the description of the summary results for all cases, we illustrate the behavior of the technique on one of the 27 forecasts (Fig. 4): P^f is a 4-km resolution, 24-h forecast of the 1-h accumulation valid at 0000 UTC 1 June 2005 by the NCEP limited area model, and P^a is the corresponding stage II precipitation analysis (Lin and Mitchell 2005). The overall structure of the main storm system in this forecast is reasonably well matched, but displaced to the west. We choose this forecast case because it has been extensively studied by others (Ahijevych et al. 2009; Wernli et al. 2009; Davis et al. 2009; Marzban et al. 2009; Mittermaier and Roberts 2010; Lakshmanan and Kain 2010; Gilleland et al. 2010). Our result for this case (right panel of Fig. 4) is in agreement with that of Lakshmanan and Kain (2010), and the shift produced by the morphing-based technique removes almost completely the location error.

Table 3 presents the correlations between the subjective scores and a number of objective scores that include ASSIM. In this table, D and D^0 are the same as before (as in Table 2), while D^c is ASSIM after correction for the location error by shifting the forecast feature, such that its center of mass is aligned with the center of mass of the verifying feature. The purpose of computing D^c in addition to D is to assess the value of using the morphing-based algorithm rather than a computationally less complex approach for the correction of the location error. The table shows that the correlation for D is higher than for the other objective scores, except for the multiplicative bias, which is a traditional measure of the amplitude error. In addition, a comparison of the first
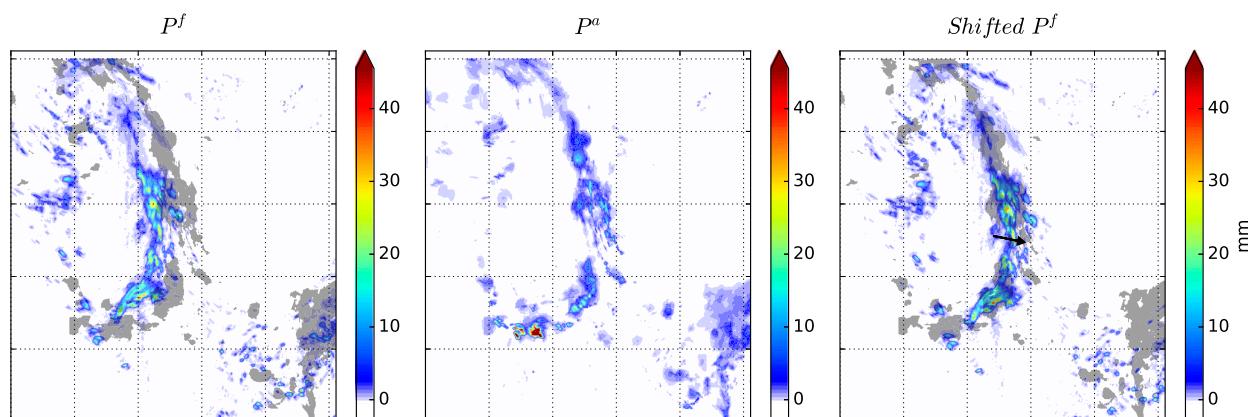

FIG. 4. Illustration of a realistic ICP example. Shown are (left) a 24-h forecast of the 1-h precipitation total by the NCEP WRF Model, (middle) the related Stage II analysis of 1-h accumulation, and (right) the shifted forecast obtained by the proposed morphing technique. The gray shading in the left and right panels shows the outline of P^a (the field in the middle panel). The black arrow in the right panel indicates the magnitude and direction of $d\mathbf{X}_{final}$ (the total shift vector).

TABLE 3. The correlations between the subjective scores and a number of objective measures for the ICP realistic cases. We use the values of the subjective score, ETS, and multiplicative bias as they were reported by Keil and Craig (2009) for our calculations. DAS is the displacement and amplitude score proposed by Keil and Craig (2009), ETS is the equitable threat score, and multiplicative bias is the amplitude error.

Measure	Correlation
D^0	0.26
D	0.47
D^c	0.32
DAS	0.26
ETS	0.41
Multiplicative bias	0.53

TABLE 4. The correlations between the subjective scores and the components of the multicomponent measures ASSIM, SAL (Wernli et al. 2009), and DAS (Keil and Craig 2009).

Measure		Type	Correlation
ASSIM before correction of location	D_{cor}^0	Structure	0.17
ASSIM after correction of location	D_{cor}	Structure	0.37
	D_{var}	Structure	0.60
	D_{amp}	Amplitude	0.56
	$|d\mathbf{X}_{final}|$	Location	0.23
DAS	AMP	Amplitude	0.25
	DIS	Location	0.05
SAL	$|S|$	Structure	−0.23
	$|A|$	Amplitude	0.47
	L	Location	0.13

two rows indicates that correcting for the location error improves the correspondence between ASSIM and the subjective scores, while a comparison of the second and third rows indicates that correcting for the location error leads to a higher correlation with the subjective scores.

Table 4 presents the correlations between the subjective scores and the components of the multicomponent measures ASSIM, structure–amplitude–location (SAL) (Wernli et al. 2009), and DAS (Keil and Craig 2009). The correlations are the highest for the variance and amplitude components of ASSIM, which suggests that these two error components reflect the human perception of forecast quality the best. Interestingly, the correlations are low for the location error component of all three measures. This result supports the argument of Wang and Bovik (2009) that the human visual system is less sensitive to small displacement errors than most objective scores.

c. MesoVICT cases

MesoVICT tier 1 is a set of multiple forecasts of six forecast cases for a range of meteorological phenomena in and around the Alps (Dorninger et al. 2013). The forecast dataset comprises 18-h-long forecasts of the hourly accumulation of the precipitation from two NWP models: the Canadian high-resolution model (CMH; 2.5-km resolution) from Environment and Climate Change Canada and the COSMO2 model (CO2; 2.2-km resolution) from MeteoSwiss. The dataset also includes Vienna Enhanced Resolution Analyses (VERA; 8-km resolution; Steinacker et al. 2000) for the verification of the forecasts. The fields of the selected models are provided as ACSII files interpolated onto the VERA grid. We use cases 2–5 to compare the forecast performance of the two models because these are the cases for which the forecasts with the two models start at the same analysis time. The total number of forecasts of the hourly accumulation is

212. For a different type of verification of the same cases, see Geiß (2015) and Gilleland (2017).

Figure 5 shows the analyses (top panel) and the forecasts started at 0600 UTC 19 July 2007 (middle and bottom panels) for four consecutive verification times (from 0800 to 1100 UTC 19 July 2017) for case 2. It also shows $d\mathbf{X}_{final}$ (black arrow) and the value of D for each forecast. While the forecasts from both models exhibit location errors that can be corrected by a northeast shift of the precipitation features, the magnitude of the location errors is clearly larger for the CMH than the CO2 model. In addition, the value of ASSIM after the correction for the location error, D, is slightly lower for the CMH model in all cases, which indicates that the amplitude and structure error is also somewhat larger for that model.

Figure 6, which shows the histogram of ASSIM values D after the correction of the location errors, also supports the conclusion that the CO2 model performs better than the CMH model in predicting the amplitude and structure of the precipitation field for the selected cases (bottom panel). For instance, ASSIM is found to be higher than $D = 0.7$ for 44% of the CO2 forecasts and only for 24% of the CMH forecasts. The figure also indicates that predicting the amplitude and structure of a precipitation event is more difficult in some cases than in others (top four panels). Interestingly, the high relative frequency at the tails of the histogram for case 2 indicates that both models produce the largest number of unusually accurate and poor forecasts of the amplitude and the structure for case 2.

4. Precipitation associated with winter storms

The results of this section are based on forecasts of the 32 U.S. winter storms that were named by The Weather Channel in the 2014/15 and 2015/16 storm seasons. While The Weather Channel does not provide a rigorous formal definition of the criteria for naming storms,
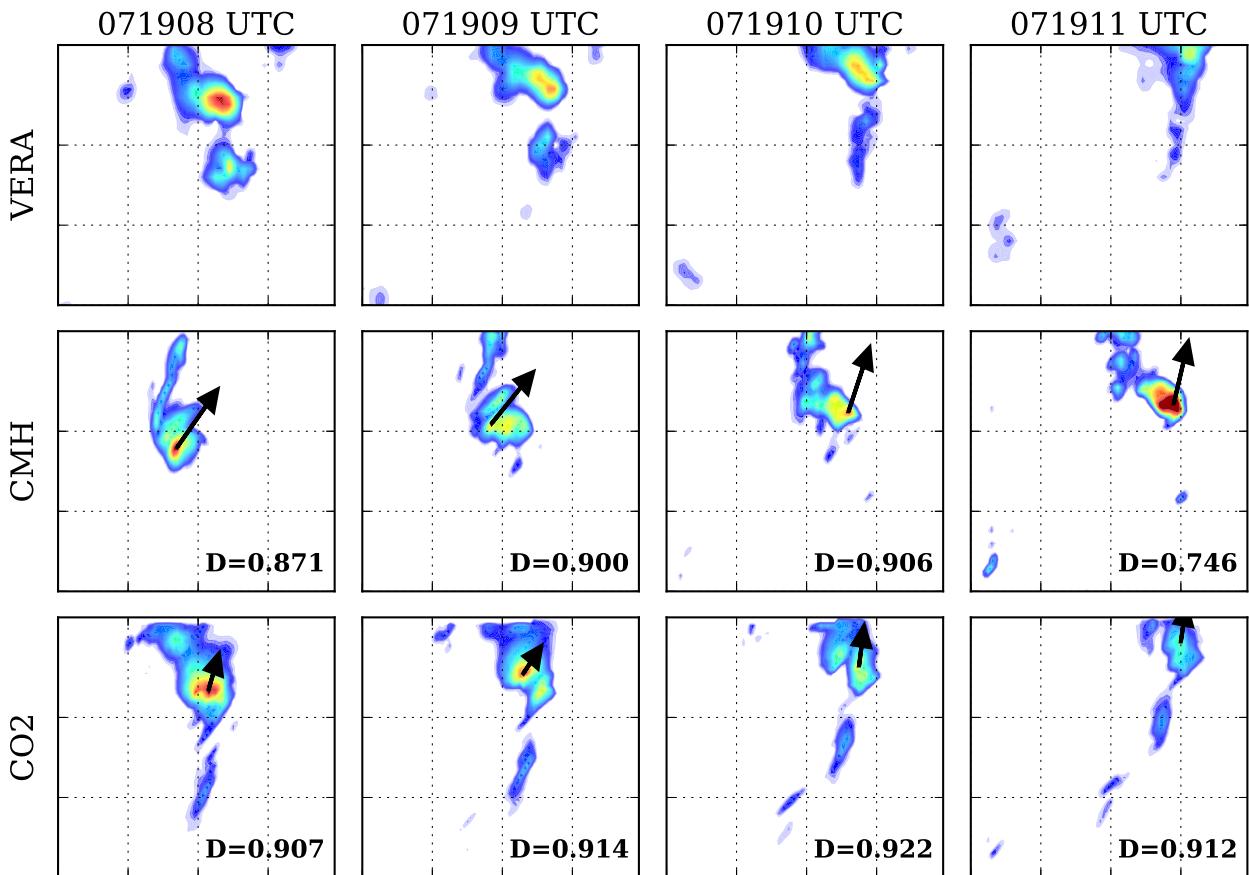
FIG. 5. An example of the verification results for a pair of forecasts from the MesoVICT tier 1 dataset. Shown are (top) the VERA analyses, (middle) the CMH forecasts, and (bottom) the CO2 forecasts for four consecutive forecast hours. Each column is a different forecast hour, with time increasing from left to right.

the named storms provide a representative collection of synoptic-scale precipitation events with potentially high socioeconomic impact. The operational forecasts are twice-daily global medium-range forecasts started at 0000 and 1200 UTC from the National Centers for Environmental Prediction (NCEP), while the verifying analyses are Stage IV precipitation analyses over land and operational 0–6-h precipitation forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF) over ocean. We homogenize the verification data from the two sources by a calibration of the ECMWF data. This calibration is done such that the amplitude and the variance of the precipitation in the ECMWF data become equal to that of the Stage IV data for the locations where data are available from both sources. At those locations, the ECMWF data are used only for the calibration. The spatial resolution of the forecast and verification dataset is $0.5° \times 0.5°$. Only those forecast cases are used for the estimation of the location error for which at least 30% of the total precipitation in the verification region is from the area

covered by Stage IV data. There are a total of 83 forecast cases that satisfy this criterion for the 32 storms.

a. Predictability of the location

This example illustrates how the proposed technique can be used to study the predictability of the location of a particular type of precipitation feature. Unlike in the earlier examples, here, we do not have a predetermined pair of features in P^f and P^a. Instead, we have a P^a that includes an observed (analyzed) precipitation feature of interest and different lead time forecasts of P^a that may or may not include a related forecast feature. Our task is to determine for each forecast P^f whether or not it includes a feature that can be considered a forecast of the forecast feature of interest in P^a. (In essence, in this application, our technique is used as an object-based verification approach.) If P^f includes a matching feature, the task also includes the computation of the location error for that feature. We obtain the curve that describes the dependence of the typical location error on the lead time by averaging
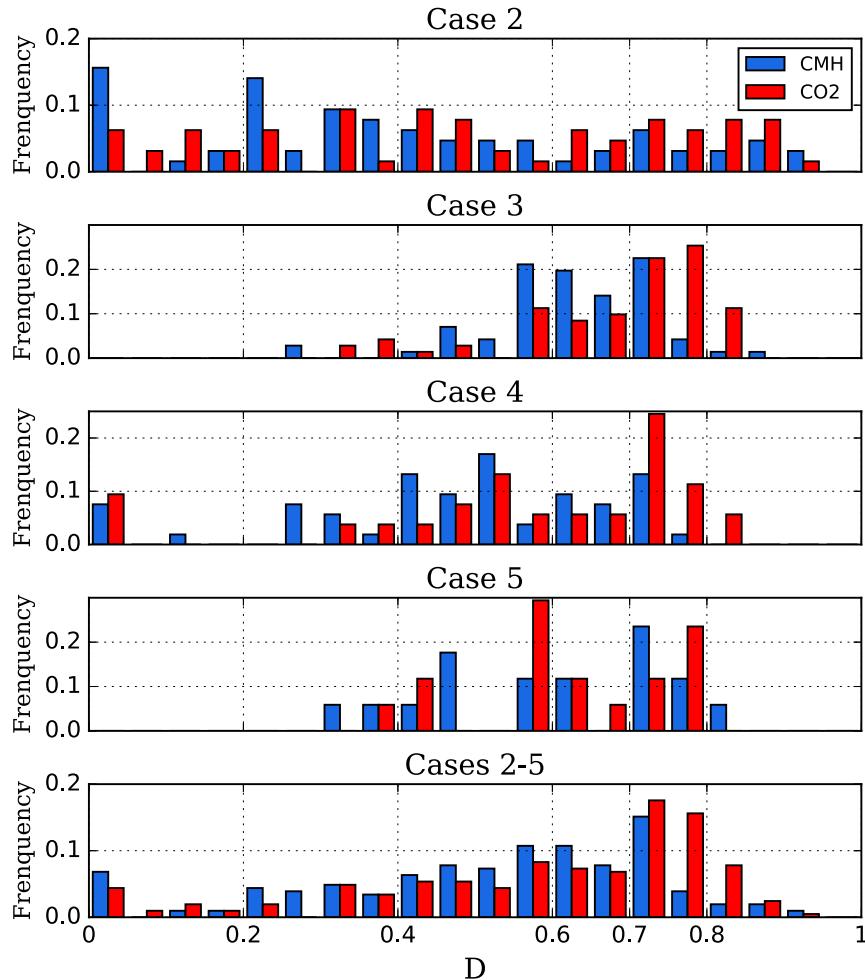
FIG. 6. Histogram of the value D of ASSIM after the correction of the location error for MesoVICT tier 1 cases 2–5. (top four panels) Histogram for the individual cases, and (bottom) histogram for the four cases combined.

the magnitude of the estimated location error for all forecasts of the same lead time.

In this application, we use ASSIM as a similarity index rather than an error measure: the morphing-based algorithm is applied to all forecasts, and a forecast storm is deemed likely to be related to the verifying storm if ASSIM, that is, the value of D defined by Eq. (6), is larger than a prescribed threshold value δ after the correction of the location error. Figure 7 shows the number of forecasts that satisfy this criteria for the different values of δ in the range $0.6 \le \delta \le 0.8$. Qualitatively, all curves in the figure behave similarly: as forecast time increases, the number of storms that satisfy the search criteria decreases until about forecast time 11 days. Under the assumption that the number of related forecast events decreases monotonically with the lead time, the saturation level of these curves (normalized by the total number of forecasts) is an estimate of

the upper bound of the probability that a forecast event is not related to the verifying event for the given value of δ. For instance, this upper bound is about 40% for $\delta = 0.6$ and 5% for $\delta = 0.8$. It could also be pushed to 0% by further increasing δ, but then an increasing number of events would be found unlikely to be related to the verifying event, even at the shortest forecast times. For example, the upper bound becomes 0% at day 10.5 for both $\delta = 0.85$ and $\delta = 0.9$ (not shown), but then 15% and 35%, respectively, of the forecast events are found unlikely to be related to the verifying event at initial time.

Remarkably, the dependence of the number of events on δ does not affect the estimate of the average location error (Fig. 8): the functional dependence of the average location error on the forecast lead time is independent of δ in the range $0.6 \le \delta \le 0.8$. This functional dependence qualitatively closely resembles the one that describes the dependence of the root-mean-square error
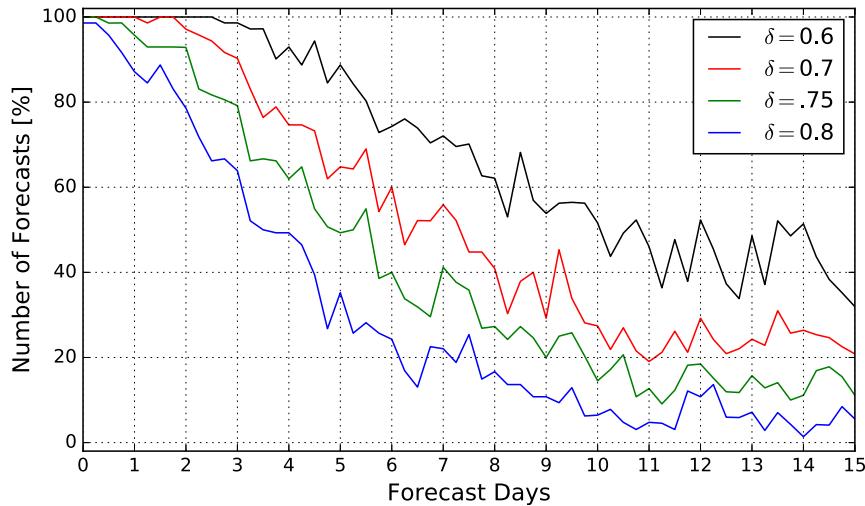
FIG. 7. Evolutions of the percentage of forecast cases in which a forecast precipitation is considered similar to the analysis measured by ASSIM. Different curves correspond to different threshold values δ of ASSIM.

of the Eulerian forecast variables (e.g., geopotential height of a specific pressure level) on the forecast lead time (e.g., Fig. 3.11 of Szunyogh 2014). The most widely used technique (e.g., Magnusson and Källén 2013; Herrera et al. 2016; Žagar et al. 2017) for the quantitative analysis of this functional dependence is the Lorenz curve analysis (Lorenz 1969, 1982; Dalcher and Kalnay 1987). The Lorenz curve analysis technique makes the assumption that the time evolution of the magnitude E of the error ($E = |\delta\mathbf{X}_{\text{final}}|$ in our case) can be described by the differential equation

$$\frac{dE}{dt} = (\alpha E + \beta)\left(1 - \frac{E}{E_{\infty}}\right), \tag{10}$$

and estimates the parameters α, β, and E_{∞} by fitting a 2nd-order polynomial to pairs of E, dE/dt values by the method of least squares. The graphical image of the fitted polynomial in the $E - dE/dt$ plane is the Lorenz curve. In addition, the quality of the fit can be verified by substituting the estimates of the parameters into the analytical solution of Eq. (10), which is

$$E(t) = E_{\infty} - \frac{E_{\infty} + \beta/\alpha}{1 + \mu(t)}, \tag{11}$$

where

$$\mu(t) = \frac{E_0 + \beta/\alpha}{E_{\infty} - E_0} \exp\left[\left(\alpha + \frac{\beta}{E_{\infty}}\right)t\right]. \tag{12}$$

In Eq. (10), E_{∞} is the saturation level of the magnitude of the error: once E reaches this value, dE/dt becomes zero,

and the error can no longer grow. The forecast time at which this happens is the predictability limit, the time by which all forecasts lose their dependence on the (unknown) true state of the atmosphere at the beginning of the forecasts. We note that for the Eulerian forecast variables, E_{∞} is equal to $\sqrt{2}$ times the root-mean-square distance between pairs of randomly selected but climatologically feasible states of the atmosphere (e.g., Leith 1974).

The predictability limit and the behavior of the three parameters of the Lorenz curve are strongly scale dependent (Žagar et al. 2017). When the technique is applied to the full forecast error, that is, not scale selectively, E_{∞} is determined by the saturation level of the errors at the synoptic and large scales. Hence, for the first few forecast days, $E \ll E_{\infty}$, and Eq. (10) can be replaced by the linear differential equation,

$$\frac{dE}{dt} = \alpha E + \beta. \tag{13}$$

The analytical solution of this differential equation is

$$E(t) = \left(E_0 + \frac{\beta}{\alpha}\right)e^{\alpha t} - \beta/\alpha, \tag{14}$$

where $E_0 = E(0)$. Thus, the linear coefficient α describes the exponential growth of the forecast errors in the forecast range for which Eq. (13) is valid. The parameter β has been conventionally interpreted as the contribution of model errors to the error growth (e.g., Dalcher and Kalnay 1987; Magnusson and Källén 2013) because according to Eq. (13), it describes error growth
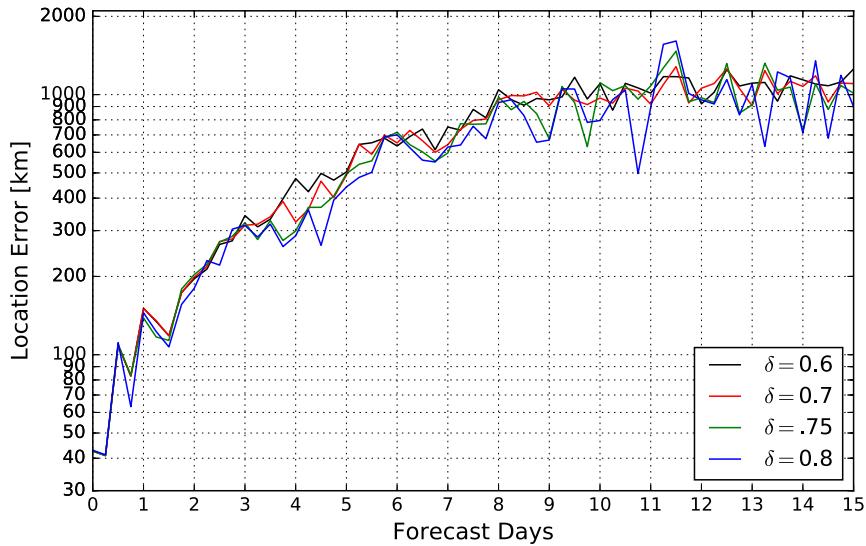
FIG. 8. Evolutions of the magnitude of location error averaged over the forecast cases selected using different threshold values δ of ASSIM.

that does not require the prior presence of errors in the forecasts.

The function fitted to the estimates of the magnitude of the location error and the related Lorenz curve for $\delta = 0.7$ are shown by solid black lines in Fig. 9. The calculation of the estimated parameters followed the procedure described in Herrera et al. (2016). To verify

that the estimates of the parameters of the fitted curve are robust, we also computed a second set of estimates of the parameters by the technique introduced by Žagar et al. (2017). Because this technique is based on estimating the parameters by fitting a curve directly in the $t - E(t)$ rather than the $E(t) - dE/dt$ plane, it eliminates the necessity of computing approximate values of dE/dt.
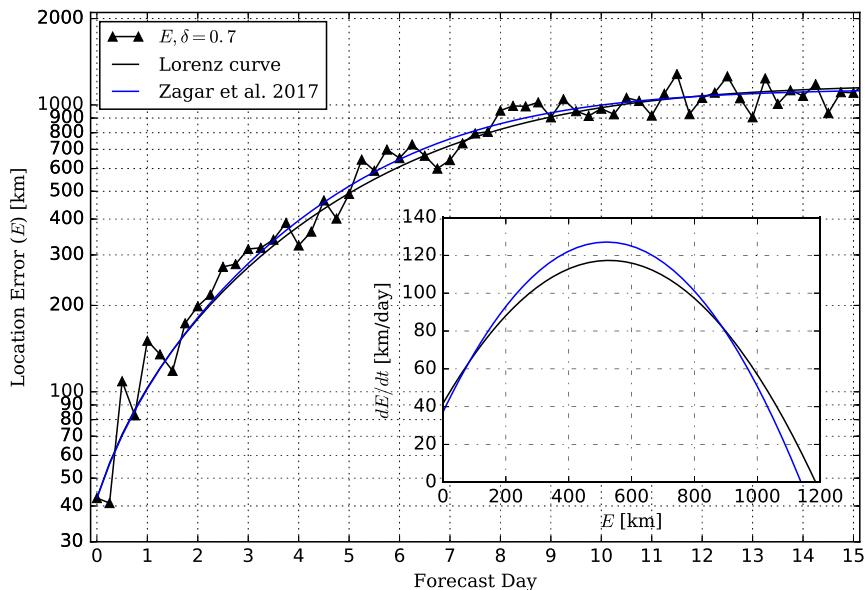

FIG. 9. The results of the Lorenz curve analysis for the location error of the precipitation events associated with the winter storms. Shown are (black triangles) the average magnitude E of the estimated location errors $\delta\mathbf{X}_{\text{final}}$ for all forecasts and the fitted curves based on Eq. (10). The inset shows the related Lorenz curves.

TABLE 5. The estimated values of the parameters of Eq. (10).

Method of estimation	α (day^{-1})	$e^{\alpha t}$ ($t = 1$ day)	β (km day^{-1})	E_∞ (km)
Herrera et al. (2016)	0.32	1.38	41.6	1186
Žagar et al. (2017)	0.38	1.46	37.1	1139

The fitted curve and the Lorenz curve for this technique are shown by solid blue lines in the figure. The actual estimates of the parameters obtained by the two approaches are presented in Table 5. While there are some differences between the estimates obtained by the two different approaches, the relative values of these differences are small. The fitted curves indicate that the average magnitude of the estimated location error is about 40 km at initial time, 110 km at day 1, 250 km at day 3, and 750 km at week 1, and it saturates at about week 2. Once the magnitude of the error reaches a saturation level, the accuracy of the numerical forecast of the location is no better than that of a randomly generated (but climatologically feasible) forecast of the location. The quantitative estimate of the saturation level (E_∞), however, should be interpreted with caution. In our study on the behavior of ensemble forecast of the same storms (Han and Szunyogh 2018), we found that beyond the week 1 forecast lead time, the forecast uncertainty of the location becomes so large that the probability that our technique cannot match the forecast feature to the verifying feature for the selected search region is nonzero. This effect leads to an underestimation of the average magnitude of the location error beyond the week 1 forecast lead time, which becomes more severe as the forecast lead time increases.

Finally, we note that the estimates $e^{\alpha t} = 1.38 - 1.46$ of the short-term error growth are similar to those

obtained by Herrera et al. (2016) for the operational deterministic forecasts of the meridional wind component at 500 hPa. Because the location error reflects errors in the prediction of the location of synoptic-scale features, this result is in agreement with the view that the dominant growing errors in the short- and medium-range global forecasts in the extratropics are errors in the prediction of the position of the synoptic-scale flow features (e.g., Snyder 1999; Oczkowski et al. 2005).

b. An example

Figures 10 and 11 show the growth of the location error for Winter Storm Cato. The main precipitation feature in these forecast cases is a precipitation band centered at approximately 35°N, 78°W (middle panels). In the 108-h forecast (Fig. 10), the precipitation band is misplaced to the east. The morphing-based technique removes the misplacement error by shifting the forecast band by about 570 km to the west, leading to an increase of ASSIM from 0.691 to 0.835. If a threshold value of $\delta = 0.8$ was used, the forecast feature would be considered likely to be related to the verifying feature. Figure 11 shows the 180-h forecast of the same precipitation event. In this case, the morphing technique shifts the narrow forecast precipitation band in an attempt to match the precipitation system of the verifying system. The value of ASSIM increases from 0.354 to 0.756 by the shift, but does not reach the threshold value $\delta = 0.8$. Also, the magnitude of the derived shift vector
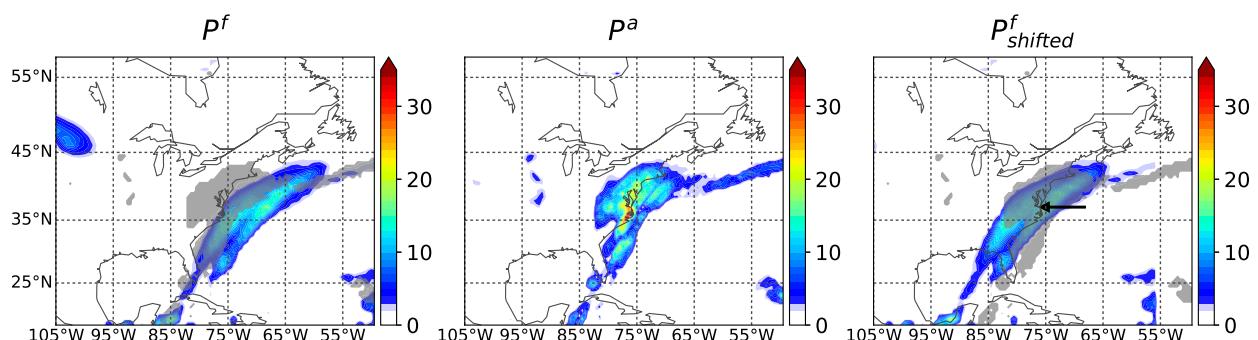

FIG. 10. Illustration of 108-h forecast of Winter Storm Cato. (left) NCEP forecast of the 6-h precipitation total at 0000 UTC 26 Nov 2014, (middle) the verifying analysis, and (right) the forecast shifted by $d\mathbf{X}_{final}$. The gray shading in the left and right panels shows the outline of P^a (the field in the middle panel). The black arrow in the right panel is the shift vector ($d\mathbf{X}_{final}$) derived out of the morphing algorithm.

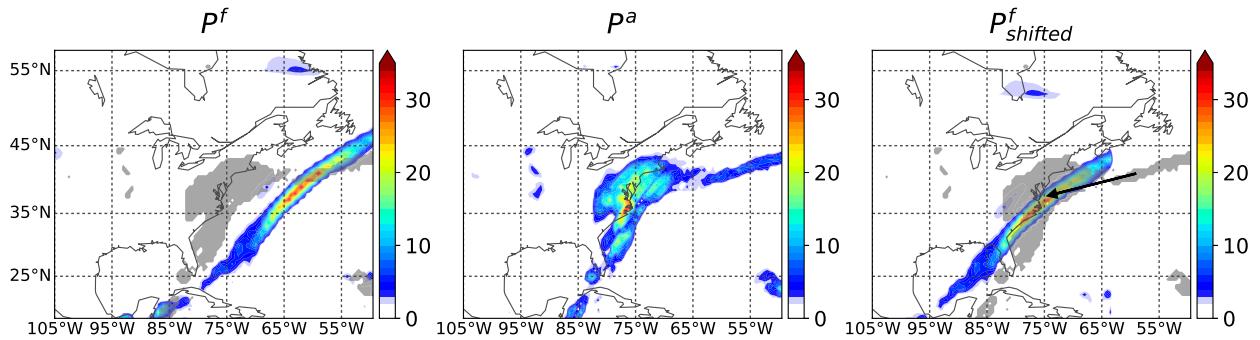FIG. 11. As in Fig. 10, but for the 180-h forecast of Winter Storm Cato.

$d\mathbf{X}$ (1270 km) is larger than the location error associated with the predictability limit (about 1140–1190 km), making it highly unlikely that there is a dynamical relationship between the forecast and the verifying feature.

5. Conclusions

In this paper, we proposed an improved morphing-based technique for the verification of precipitation forecasts for which the location error is well defined and can be described by a spatial shift (e.g., precipitation events related to the passage of an extratropical or tropical cyclone or a form of organized convection). This technique describes the estimate of the forecast error by a multivariate measure. One component of the measure characterizes the location error, while the other component, which can be broken up into multiple components, measures the amplitude and structure error. We introduced a new similarity index called ASSIM to measure this component of the error. ASSIM is an adaptation of the Wang–Bovik image quality index, and in addition to using it as an error measure, we also use it to define the convergence criteria for the morphing algorithm.

We illustrated the behavior of the new morphing technique and ASSIM through applications to idealized and realistic forecast cases. While the results of these applications are highly encouraging, it is important to note that in the investigated cases, the verifying precipitation features and the verification domains around them were selected manually. An automation of the technique for the verification of operational real-time forecasts would require further research into the selection of the events and the verification domain. In addition, because the technique considers the precipitation in the verification domain part of a single precipitation system, it is not suitable to verify forecasts in a preselected fixed domain if precipitation in different parts

of the domain may not be related (e.g., convective precipitation with no larger-scale organization).

In one of our examples, we studied the predictability of the location of U.S. winter storms. It was found that the functional dependence of the average magnitude of the location error on the forecast lead time is qualitatively similar to that for the other conventional forecast variables (e.g., geopotential height and temperature). Quantitatively, the average magnitude of the estimated location error is about 40 km at initial time, 110 km at day 1, 250 km at day 3, and 750 km at week 1, and it eventually saturates at about week 2.

Finally, it should be noted that the proposed technique could also be used for the verification of other scalar forecast parameters whose fields have sharp gradients (e.g., potential vorticity). For such parameters, the proposed approach could provide a more nuanced picture about the evolution of the errors than an approach based on the root-mean-square error, which would indicate a sudden and complete loss of predictability of the parameter after the development of a relatively small location error. For instance, in an ongoing research project, we use the technique to assess the effect of mesoscale ocean eddies on the atmospheric state variables along the Pacific storm track.

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, https://doi.org/10.1175/2009WAF2222298.1.

Dalcher, A., and E. Kalnay, 1987: Error growth and predictability in operational ECMWF forecasts. *Tellus*, **39A**, 474–491, https://doi.org/10.3402/tellusa.v39i5.11774.

Davis, C. A., B. G. Brown, R. Bullock, and J. Halley-Gotway, 2009: The method for object-based diagnostic evaluation (MODE) applied to numerical forecasts from the 2005 NSSL/SPC Spring Program. *Wea. Forecasting*, **24**, 1252–1267, https://doi.org/10.1175/2009WAF2222241.1.

Dorninger, M., M. P. Mittermaier, E. Gilleland, E. E. Ebert, B. G. Brown, and L. J. Wilson, 2013: MesoVICT: Mesoscale verification inter-comparison over complex terrain. NCAR Tech. Note NCAR/TN-505+STR, 23 pp., https://doi.org/10.5065/D6416V21.

Geiß, S., 2015: Comparison of spatial verification methods. B.A. thesis, Munich Meteorological Institute, Ludwig-Maximilians-University, 43 pp.

Gilleland, E., 2011: Spatial forecast verification: Baddeley's delta metric applied to the ICP test cases. *Wea. Forecasting*, **26**, 409–415, https://doi.org/10.1175/WAF-D-10-05061.1.

——, 2017: A new characterization within the spatial verification framework for false alarms, misses, and overall patterns. *Wea. Forecasting*, **32**, 187–198, https://doi.org/10.1175/WAF-D-16-0134.1.

——, J. Lindström, and F. Lindgren, 2010: Analyzing the image warp forecast verification method on precipitation fields from the ICP. *Wea. Forecasting*, **25**, 1249–1262, https://doi.org/10.1175/2010WAF2222365.1.

Han, F., and I. Szunyogh, 2016: A morphing-based technique for the verification of precipitation forecasts. *Mon. Wea. Rev.*, **144**, 295–313, https://doi.org/10.1175/MWR-D-15-0172.1.

——, and ——, 2018: How well can an ensemble predict the uncertainty in the location of winter storm precipitation? *Tellus*, **70A**, 1440870, https://doi.org/10.1080/16000870.2018.1440870.

Herrera, M. A., I. Szunyogh, and J. Tribbia, 2016: Forecast uncertainty dynamics in the THORPEX Interactive Grand Global Ensemble (TIGGE). *Mon. Wea. Rev.*, **144**, 2739–2766, https://doi.org/10.1175/MWR-D-15-0293.1.

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, https://doi.org/10.1175/MWR3457.1.

——, and ——, 2009: A displacement and amplitude score employing an optical flow technique. *Wea. Forecasting*, **24**, 1297–1308, https://doi.org/10.1175/2009WAF2222247.1.

Lack, S. A., G. L. Limpert, and N. I. Fox, 2010: An object-oriented multiscale verification scheme. *Wea. Forecasting*, **25**, 79–92, https://doi.org/10.1175/2009WAF2222245.1.

Lakshmanan, V., and J. S. Kain, 2010: A Gaussian mixture model approach to forecast verification. *Wea. Forecasting*, **25**, 908–920, https://doi.org/10.1175/2010WAF2222355.1.

Leith, C., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102**, 409–418, https://doi.org/10.1175/1520-0493(1974)102<0409:TSOMCF>2.0.CO;2.

Lin, Y., and K. E. Mitchell, 2005: The NCEP Stage II/IV hourly precipitation analyses: Development and applications. *19th Conf. on Hydrology,* San Diego, CA, Amer. Meteor. Soc., 1.2, https://ams.confex.com/ams/Annual2005/techprogram/paper_83847.htm.

Lorenz, E. N., 1969: Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmos. Sci.*, **26**, 636–646, https://doi.org/10.1175/1520-0469(1969)26<636:APARBN>2.0.CO;2.

——, 1982: Atmospheric predictability experiments with a large numerical model. *Tellus*, **34**, 505–513, https://doi.org/10.3402/tellusa.v34i6.10836.

Magnusson, L., and E. Källén, 2013: Factors influencing skill improvements in the ECMWF forecasting system. *Mon. Wea. Rev.*, **141**, 3142–3153, https://doi.org/10.1175/MWR-D-12-00318.1.

Marzban, C., S. Sandgathe, H. Lyons, and N. Lederer, 2009: Three spatial verification techniques: Cluster analysis, variogram, and optical flow. *Wea. Forecasting*, **24**, 1457–1471, https://doi.org/10.1175/2009WAF2222261.1.

Mittermaier, M., and N. Roberts, 2010: Intercomparison of spatial forecast verification methods: Identifying skillful spatial scales using the fractions skill score. *Wea. Forecasting*, **25**, 343–354, https://doi.org/10.1175/2009WAF2222260.1.

Mo, R., C. Ye, and P. H. Whitfield, 2014: Application potential of four nontraditional similarity metrics in hydrometeorology. *J. Hydrometeor.*, **15**, 1862–1880, https://doi.org/10.1175/JHM-D-13-0140.1.

Oczkowski, M., I. Szunyogh, and D. Patil, 2005: Mechanisms for the development of locally low-dimensional atmospheric dynamics. *J. Atmos. Sci.*, **62**, 1135–1156, https://doi.org/10.1175/JAS3403.1.

Snyder, C., 1999: Error growth in flows with finite-amplitude waves or coherent structures. *J. Atmos. Sci.*, **56**, 500–506, https://doi.org/10.1175/1520-0469(1999)056<0500:EGIFWF>2.0.CO;2.

Steinacker, R., C. Häberli, and W. Pöttschacher, 2000: A transparent method for the analysis and quality evaluation of irregularly distributed and noisy observational data. *Mon. Wea. Rev.*, **128**, 2303–2316, https://doi.org/10.1175/1520-0493(2000)128<2303:ATMFTA>2.0.CO;2.

Szunyogh, I., 2014: *Applicable Atmospheric Dynamics: Techniques for the Exploration of Atmospheric Dynamics.* World Scientific, 608 pp.

Venugopal, V., S. Basu, and E. Foufoula-Georgiou, 2005: A new metric for comparing precipitation patterns with an application to ensemble forecasts. *J. Geophys. Res.*, **110**, D08111, https://doi.org/10.1029/2004JD005395.

Wang, Z., and A. C. Bovik, 2002: A universal image quality index. *IEEE Signal Process. Lett.*, **9**, 81–84, https://doi.org/10.1109/97.995823.

——, and ——, 2009: Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE Signal Process. Mag.*, **26**, 98–117, https://doi.org/10.1109/MSP.2008.930649.

——, ——, H. R. Sheikh, and E. P. Simoncelli, 2004: Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.*, **13**, 600–612, https://doi.org/10.1109/TIP.2003.819861.

Wernli, H., M. Paulat, M. Hagen, and C. Frei, 2008: SAL—A novel quality measure for the verification of quantitative precipitation forecasts. *Mon. Wea. Rev.*, **136**, 4470–4487, https://doi.org/10.1175/2008MWR2415.1.

——, C. Hofmann, and M. Zimmer, 2009: Spatial forecast verification methods intercomparison project: Application of the SAL technique. *Wea. Forecasting*, **24**, 1472–1484, https://doi.org/10.1175/2009WAF2222271.1.

Žagar, N., M. Horvat, Ž. Zaplotnik, and L. Magnusson, 2017: Scale-dependent estimates of the growth of forecast uncertainties in a global prediction system. *Tellus*, **69A**, 1287492, https://doi.org/10.1080/16000870.2017.1287492.

Zhu, M., V. Lakshmanan, P. Zhang, Y. Hong, K. Cheng, and S. Chen, 2011: Spatial verification using a true metric. *Atmos. Res.*, **102**, 408–419, https://doi.org/10.1016/j.atmosres.2011.09.004.