

## Optical Flow for Verification

CAREN MARZBAN

*Applied Physics Laboratory and Department of Statistics, University of Washington, Seattle, Washington*

SCOTT SANDGATHE

*Applied Physics Laboratory, University of Washington, Seattle, Washington, and College of Oceanic and Atmospheric Sciences, Oregon State University, Corvallis, Oregon*

(Manuscript received 25 August 2009, in final form 11 June 2010)

### ABSTRACT

Modern numerical weather prediction (NWP) models produce forecasts that are gridded spatial fields. Digital images can also be viewed as gridded spatial fields, and as such, techniques from image analysis can be employed to address the problem of verification of NWP forecasts. One technique for estimating how images change temporally is called optical flow, where it is assumed that temporal changes in images (e.g., in a video) can be represented as a fluid flowing in some manner. Multiple realizations of the general idea have already been employed in verification problems as well as in data assimilation. Here, a specific formulation of optical flow, called Lucas–Kanade, is reviewed and generalized as a tool for estimating three components of forecast error: intensity and two components of displacement, direction and distance. The method is illustrated first on simulated data, and then on a 418-day series of 24-h forecasts of sea level pressure from one member [the Global Forecast System (GFS)–fifth-generation Pennsylvania State University–National Center for Atmospheric Research Mesoscale Model (MM5)] of the University of Washington’s Mesoscale Ensemble system. The simulation study confirms (and quantifies) the expectation that the method correctly assesses forecast errors. The method is also applied to a real dataset consisting of 418 twenty-four-hour forecasts spanning 2 April 2008–2 November 2009, demonstrating its value for analyzing NWP model performance. Results reveal a significant intensity bias in the subtropics, especially in the southern California region. They also expose a systematic east-northeast or downstream bias of approximately 50 km over land, possibly due to the treatment of terrain in the coarse-resolution model.

### 1. Introduction

A numerical weather prediction (NWP) model produces spatial fields of forecast parameters generally as a complex array of spatial features placed on a regular two-dimensional grid. Although observations are gathered at irregular points and times, they are dynamically assimilated to a uniform grid as well, producing a two-dimensional analysis field of a parameter. The forecasts may then be verified by comparing the two complex gridded fields. It has been shown, however, that some of the simpler notions of comparison produce misleading results at finer grid resolutions. For example, taking a difference between corresponding grid points from the two fields

penalizes the forecasts twice (but for the same error) if a forecast feature such as a frontal zone or wind maximum does not overlap or only partially overlaps the observed feature (Brown et al. 2004; Davis et al. 2006). For such reasons, a number of verification techniques have been developed, wherein the spatial structure of the two fields is taken into account. A classification and a review of these techniques appear in Casati et al. (2008). The references in that article are quite extensive, but the following works are also relevant: Gebremichael et al. (2004) and Germann and Joss (2001).

Hoffman et al. (1995) emphasize ways in which one field (e.g., forecast) can be transformed into the other (e.g., observed). They examine a specific type of transformation, called distortion, which is a combination of spatial displacement and changes in intensity. Any errors that cannot be accounted for by distortion are considered to be residual errors (e.g., rotation or shape). Variations on this theme have been put forth, for nowcasting, verification,

---

*Corresponding author address:* Caren Marzban, Dept. of Statistics, Box 354322, University of Washington, Seattle, WA 98195-4322.

E-mail: marzban@stat.washington.edu

DOI: 10.1175/2010WAF2222351.1

© 2010 American Meteorological Society

waf2222351

and/or assimilation, by Douglas (2002), Brill (2002), Du et al. (2000), Hoffman and Grassotti (1996), Nehrkorn et al. (2003), Brewster (2003), Germann and Zawadzki (2002, 2004), and Reilly et al. (2003). Some of these works are inspired by the ability of the approach to assess the distortion error itself, because it can offer ways of possibly correcting the forecasts. Hodges (1994, 1995, 1996, 1998, 1999), too, has examined computer vision algorithms for tracking meteorological features.

In image-processing circles, estimating these transformations is called optical flow (Trucco and Verri 1998). Two meteorology papers that employ optical flow explicitly are Bowler et al. (2004) and Keil and Craig (2007). The former employs ideas from optical flow, not for verification, but for the purpose of producing nowcasts. The latter uses optical flow concepts for verification. Specifically, Keil and Craig allow for an arbitrary match between pixels in one field and pixels in the other field, and then identify the specific match that minimizes the difference (in a least squares sense) between the two images. Additionally, they embed this scheme within a pyramidal image-matching scheme, wherein the former is repeated over successively finer scales.

As indicated above, the notion of an optical flow is broad, and can include any method for mapping one image to another. Many variants take a purely algorithmic or nonparametric approach (e.g., Keil and Craig 2007), capable of modeling a wide range of flow patterns. But there is one method whose simplicity allows for even analytic solutions. The method is called the Lucas–Kanade (LK) method, and it appeared early in the development of the field (Lucas and Kanade 1981). In spite of the development of more refined methods, it has repeatedly proven itself in a wide range of applications (Baker and Matthews 2004). The underlying simplicity of the approach, however, remains as the main reason for its popularity. The next section provides the exact formulation of the optical flow equations and the LK approach to solving them, thereby rendering the underlying assumptions and limitations more transparent. It also suggests possible generalizations.

The original LK formulation assumes that the intensity of the image is constant in time, because it was designed to estimate only motion (i.e., displacement). However, meteorological features develop as well as move and, so, in this paper the original formulation is generalized to allow for the estimation of displacement and intensity errors. Also, the original formulation was linear (see next section); here, it has been generalized into a nonlinear formulation in order to provide more accurate estimates of the errors.

The explicit and transparent nature of the LK approach (both the original version and the generalization

developed here) is the reason it is adopted in the current work. This particular choice of the method for solving the equations of optical flow is also what sets this work apart from that of Hoffman et al. (1995) and Keil and Craig (2007). Although further details of the LK approach are presented below, it is a differential method (i.e., it calls for spatial derivatives to be estimated) and local (i.e., assumes that the optical flow field is locally constant). This is in contrast to the methods of Hoffman et al. (1995) and Keil and Craig (2007), where these constraints are not explicitly assumed. One benefit of the proposed method is that one can then refer to standard statistical methods and theorems for solving the problem and explaining/understanding the results; see section 7.

Finally, a note about the continuous versus discrete nature of fields is in order. The above-mentioned optical-flow-based articles deal with a variety of meteorological variables, including temperature, sea level pressure, precipitation, and reflectivity. The differential nature of the LK method calls for the existence of a well-defined derivative, and so it is more suited for continuous fields. However, its local nature allows for the weaker constraint of a piece-wise continuous derivative. In other words, as long as derivatives are computable within regions of the field, then the method still works. As such, although the method will work for mixed discrete-continuous fields, this paper deals only with continuous fields, specifically sea level pressure.

The main goals of this paper are to 1) revisit the optical flow approach within a sufficiently simple framework (i.e., Lucas–Kanade) capable of better elucidating its inner working and 2) generalize it to allow for a diagnostic decomposition of the forecast errors into intensity and displacement errors. To that end, a synthetic/simulated example is studied first, and the approach is then applied to realistic forecasts of sea level pressure from a member of the University of Washington Mesoscale Ensemble system; the member is the fifth-generation Pennsylvania State University–National Center for Atmospheric Research Mesoscale Model (MM5) initialized by the National Weather Service’s Global Forecast System (GFS).

## 2. Optical flow

The estimation of displacement fields has been well studied in image analysis circles (Chan 1993; Lim and Ho 1998). A subset of displacement estimation techniques assumes that the displacements are small; the resulting displacements are then called optical flow (OF). As such, technically, an OF field is an approximation to a displacement field. The multitude of techniques for estimating the displacement field can be classified according to their

emphasis on the underlying assumptions. At one extreme, the methods are purely algorithmic, attempting to find any map that relates one field to another. At the other extreme, the desired map is constrained through explicit/analytic constraints imposed on the OF. The explicit nature of the constraints renders the methods more transparent. In the latter group, two commonly used techniques for the estimation of the OF field are the Horn–Schunck (Horn and Schunck 1981) and LK (Lucas and Kanade 1981) methods. Both are described below because they provide different ways of viewing OF, but only the latter method is employed and generalized in the current application. As mentioned in the introduction, many of the techniques in the verification of spatial fields, as well as in data assimilation, can be considered as optical flow techniques, although they are often not known by that name.

The equations of OF are designed to estimate motion and, so, time plays a central role. Specifically, one uses multiple images at different times in order to derive the OF. The LK and Horn–Schunck methods require only two images, and for that reason are ideal for use in verification. Within the verification context, the images at the two times correspond to the two fields: observed/analyzed and forecast. Although in the following equations reference is made to time, it is to be understood that the earlier image refers to the forecast field, and the later image corresponds to the observed field. In other words, in this paper, the OF field maps the forecast field to the observed field. The inverse map is also useful because it assesses a different facet of forecast quality, but it will not be examined here. Keil and Craig (2009) describe a performance/summary measure that in fact utilizes both OF fields: the forward and the inverse maps. Also, since OF was developed for image processing, its description generally refers to pixels; here, pixels and grid points are used interchangeably.

#### a. Lucas–Kanade (LK)

Consider the intensity of an image at pixel  $(x, y)$  and at time  $t$ :  $I(x, y, t)$ . At a time  $t + dt$ , the pixels change their intensities to  $I(x, y, t + dt)$ . The basic assumption of the traditional OF approach is that the image (or field) intensity does not change appreciably from  $t$  to  $t + dt$ . So, the intensity at  $(x, y)$ , at time  $t$ , is approximately equal to that at a shifted point  $(x + dx, y + dy)$ , at time  $t + dt$ :

$$I(x, y, t) \simeq I(x + dx, y + dy, t + dt). \quad (1)$$

Within a verification context, denoting the observed and forecast fields as  $I_o$  and  $I_f$ , respectively, this equation translates to

$$I_o(x, y) \simeq I_f(x + dx, y + dy). \quad (2)$$

Approximating the right-hand side with a Taylor series expansion yields

$$I_o(x, y) \simeq I_f(x, y) + \frac{\partial I_f}{\partial x} dx + \frac{\partial I_f}{\partial y} dy. \quad (3)$$

Higher-derivative terms are assumed to be negligible in the original LK formulation. The collection of all pairs/vectors  $(dx, dy)$ , one per pixel, is called the OF field. It is these quantities in Eq. (3) that are to be estimated from two images and their spatial derivatives.

All of the derivatives in Eq. (3) can be approximated with finite differences computed from either image.<sup>1</sup> However, Eq. (3) provides only one equation with two unknowns  $(dx, dy)$ . Horn and Schunck (1981) propose a continuity constraint reflecting the assumption that nearby pixels have similar displacements. They show that one can set up a constrained optimization problem whose solutions are unique and physically reasonable. Alternatively, Lucas and Kanade (1981) assume that the OF field is locally constant. In other words, it is assumed that all of the OF vectors in a window of some size  $(W)$  are all approximately equal. But then, that leads to an overdetermined system of equations for  $dx$  and  $dy$ , because each pixel in the window contributes one equation. In LK one finds the least squares solution to this overdetermined system. In other words, for a given window, one computes a single OF vector that on average (in a least squares sense) solves Eq. (3). This vector is assigned to the pixel at the center of the window, and the procedure is repeated for similar windows centered on all other pixels in the field. The result is an OF field of vectors, one per grid point, that displays the displacement error mapping the forecast to the observed field.

Note that Eq. (3) is linear in the parameters  $dx$  and  $dy$ . As such, the least squares estimates can be computed analytically and uniquely. In the next section, two generalizations are made, one of which renders the analogous equation nonlinear.

#### b. Generalizations

As mentioned previously, one of the generalizations of LK developed here is the extension of the LK to the situation where the intensity of the field is not assumed to be constant. The specific model examined here allows for the intensity to change according to

$$I_o(x, y) \simeq I_f(x + dx, y + dy) + A(x, y). \quad (4)$$

<sup>1</sup> Experiments with data examined here suggest that higher-order differences generally do not affect the results.

The parameters to be estimated are now  $dx$ ,  $dy$ , and  $A(x, y)$ , which represents additive errors in intensity. Multiplicative errors are addressed in the discussion section. The special case with  $A(x, y) = 0$ , for all  $(x, y)$ , corresponds to the traditional OF model of Eq. (3). The  $dx$ ,  $dy$  parameters may be transformed to polar coordinates, in which case the radial and angular components represent displacement and angular errors, respectively.

The Taylor series expansion of  $I_f(x + dx, y + dy)$  leads to

$$I_o(x, y) \simeq A(x, y) + I_f(x, y) + \frac{\partial I_f}{\partial x} dx + \frac{\partial I_f}{\partial y} dy + \frac{\partial^2 I_f}{\partial x^2} dx^2 + \frac{\partial^2 I_f}{\partial y^2} dy^2 + \frac{\partial^2 I_f}{\partial x \partial y} dx dy + \dots \quad (5)$$

As in the traditional LK approach, neglecting the second and higher derivatives leads to a linear least squares problem. Note that estimating the parameters  $A(x, y)$ ,  $dx$ ,  $dy$  is equivalent to solving a regression problem with  $I_o$  as the target (or response) variable and  $I_f$ ,  $\partial I_f / \partial x$ , and  $\partial I_f / \partial y$  as the covariates (or predictors). Then, the regression coefficients are the parameters  $dx$ ,  $dy$ , and the  $y$  intercept in the regression equation represents the intensity error  $A(x, y)$ . As described in section 7, this regression interpretation/model for OF is useful because it aids in explaining some of the results, and can also preclude some pitfalls.

The inclusion of the second derivatives in Eq. (5) renders the least squares equations (not shown) truly nonlinear in the parameters, calling for a nonlinear optimization approach.<sup>2</sup> The benefit of including higher derivatives in the Taylor expansion is a more accurate estimation of the true forecast error. In general, whether or not the benefit of the higher accuracy outweighs the additional computational cost depends on the complexity/smoothness of the fields. In this work, both the linear and the nonlinear approximations are examined. Simulated data are employed to compare the two. In the case of sea level pressure data, the computational cost was found to be nonprohibitive, and so the nonlinear model is employed to fit the realistic data.

The LK approach, as well as the generalizations developed here, involve a single quantity that must be specified—the size  $W$  of a window across which the OF

field is assumed to be constant. This quantity is discussed at length in section 5a.

AU8

### 3. Method

For didactic reasons, the first dataset examined is simulated. The simulated forecast field  $I_f(x, y)$  is set to be a bivariate Gaussian over a  $50 \times 50$  grid, with a fixed mean, no correlation between the two coordinates  $(x, y)$ , and a standard deviation  $S$ . The quantity  $S$  effectively measures the spatial extent of an “object” in the forecast field. A small value corresponds to a spatially contained feature in the field, while a large value is associated with a field that is mostly constant across the entire image. A range of  $S$  values is examined, namely  $S = 7, 9, 11$ , etc. For the specific simulation set up here, values smaller than 7 lead to poorly estimated derivatives, while values larger than 11 lead to similar results because then the field is nearly constant. For this reason, only the results for the extreme values (i.e., 7 and 11) are presented.

The simulated observed field is the same as the forecast field, except that it is incrementally shifted along the  $x$  and  $y$  directions from 1 to 10 grid lengths. Additionally, the intensity across the entire grid of the forecast differs from that of the observed by an additive component that amounts to 20% of the maximum intensity of the forecast field. Therefore, for example, when the amount of shift is 1 grid length in the  $x$  and  $y$  directions, then the true value of the magnitude of displacement error is  $\sqrt{2}$ , and the true value of the angular error is  $45^\circ$ . The true value of the intensity error varies with the parameters of the Gaussian, but it is always 20% of the maximum intensity.

The quantity  $S$  is a parameter of the simulation, but the OF method itself has a parameter as well, namely  $W$ . Here, it is varied over a wide range of values (from 5 to 41). In the current formulation,  $W < 5$  values are not allowed because second derivatives require a minimum of five grid points to compute. Values larger than about 11 produce results that are mostly the same as those of  $W = 11$ . For these reasons, only the results for  $W = 5$  and  $W = 11$  are shown.

Recall that the proposed OF approach involves three parameters that must be estimated: intensity, displacement, and angular errors. Each of these parameters is a spatial field and, so, requires an image or a contour plot to display. Producing all of these figures for all the parameters of the simulation is impractical. Instead, only summary measures are computed to capture the accuracy of the error estimates. Given that the true values of the errors are known (for the simulated data), the summary measure examined here is the proportion of (or relative) error:

<sup>2</sup> Here, the algorithm used is a quasi-Newton variable metric method known as the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm (Press et al. 1999); this is one of the methods invoked in an R function called `optim()`; see R Development Core Team (2008).

$$\frac{\text{Median Estimated Error} - \text{True Error}}{\text{True Error}}. \quad (6)$$

For example, for the displacement error the quantity computed is

$$\text{PropDisErr} = \frac{\text{Median Estimated Displacement Error} - \sqrt{2}}{\sqrt{2}}. \quad (7)$$

The analogous quantities for intensity and angular error are computed similarly and are denoted PropIntErr and PropAngErr, respectively.<sup>3</sup> The dependence of these quantities on the relative shift between the forecast and observed fields is explored below.

A useful quantity to examine is the histogram of the estimates of the three errors. For instance, a peak in the histogram of the angles of the OF vectors would suggest a preponderance of vectors all pointing in a given direction. Similarly, a peak in the histogram of the magnitude of the OF vectors would imply that there is a systematic displacement relating the two fields. Therefore, in this paper, each of the three error fields associated with the three components of error is accompanied by a histogram that shows the distribution of the corresponding errors across the entire image.

Although such univariate histograms convey useful information, it is more useful to examine the joint histogram of the three quantities. This quantity is difficult to display in two dimensions, and so only the joint histogram of two of the three components is produced here. Again, considering the magnitude and angle, in a hypothetical situation where there is a peak in the histogram of directions, it is the magnitude of the vectors near this peak that would also have an important interpretation, more so than the magnitude of all vectors across the entire OF field. The joint histogram of the magnitude and direction of the OF vectors summarizes the OF field in a manner that reflects the coherence between the magnitude and direction.

#### 4. Results: Simulated data

**F1** Figure 1 contains multiple panels that together display all the various components of forecast error according to the OF approach. The underlying dataset is one of the

<sup>3</sup> As shown in the next section, the distribution of the errors is generally skewed. For that reason, the median (not the mean) is taken to represent the center of the distribution. The mode is also computed, but it is not clear if the extra effort involved with estimating a mode is justified by the increased accuracy.

above-mentioned simulations. Specifically, the forecast field is a Gaussian with a standard deviation (i.e., width) of  $S = 11$ , centered at coordinates (10, 10) on a  $50 \times 50$  grid. To obtain the observed field, the forecast field is shifted by one grid length in the  $x$  and  $y$  directions, and the number 60 (i.e., 20% of the maximum forecast field) is added to it.

The top panel in Fig. 1 shows the forecast and observed fields as contour plots: red for forecast and blue for observed. Also shown are the OF vectors as estimated according to the nonlinear scheme described above, with  $W = 9$ . To avoid clutter, every fourth vector is shown. Visually, the OF field confirms what is known about the relative position of the observed and forecast fields, namely that one is shifted with respect to the other in the  $45^\circ$  direction (measured counterclockwise from the  $x$  axis). An explanation of why the OF field appears distorted is given in section 7; it is related to the flatness of the Gaussian far from its center.

The remaining panels show the three components of the forecast error and their distributions. The spatial distribution of the errors is shown as both a grayscale image as well as a contour plot superimposed on the image.<sup>4</sup> The intensity error is mostly constant across the field, as expected. The corresponding histogram shows the distribution of these errors, and the true value of 60 (as marked by the vertical red line). Although these intensity errors vary from 54 to 64, the bulk of the histogram is around 60 (i.e., the true value), and the mode of the histogram is at 60. The displacement errors show a similar spatial pattern, and the corresponding histogram shows that they are correctly clustered around the true value of 1.414 (i.e.,  $\sqrt{2}$ ). The angular errors show a different spatial pattern, but the estimates are tightly clustered around  $45^\circ$ , again the correct value. In short, the nonlinear scheme for estimating the three error parameters produces correct results, namely relatively smooth spatial fields, with the magnitude of the respective errors clustered at approximately the correct/true values. Section 7 explains why the estimates are not exactly equal to the true values.

Figure 2 examines the accuracy of the estimates as a function of the shift between the forecast and observed fields, for two different object sizes ( $S = 7, 11$ ) and window sizes ( $W = 5, 11$ ). Consider the three panels for  $W = 5$  and  $S = 7$  (top, left). The intensity error (PropIntErr) increases with shift, but that of the nonlinear model (squares) increases much slower than that of the linear model (circles). Even with a shift as large as 10 grid lengths, the proportion

<sup>4</sup> In the grayscale images, white (black) corresponds to a minimum (maximum) value across the image.

**F2**

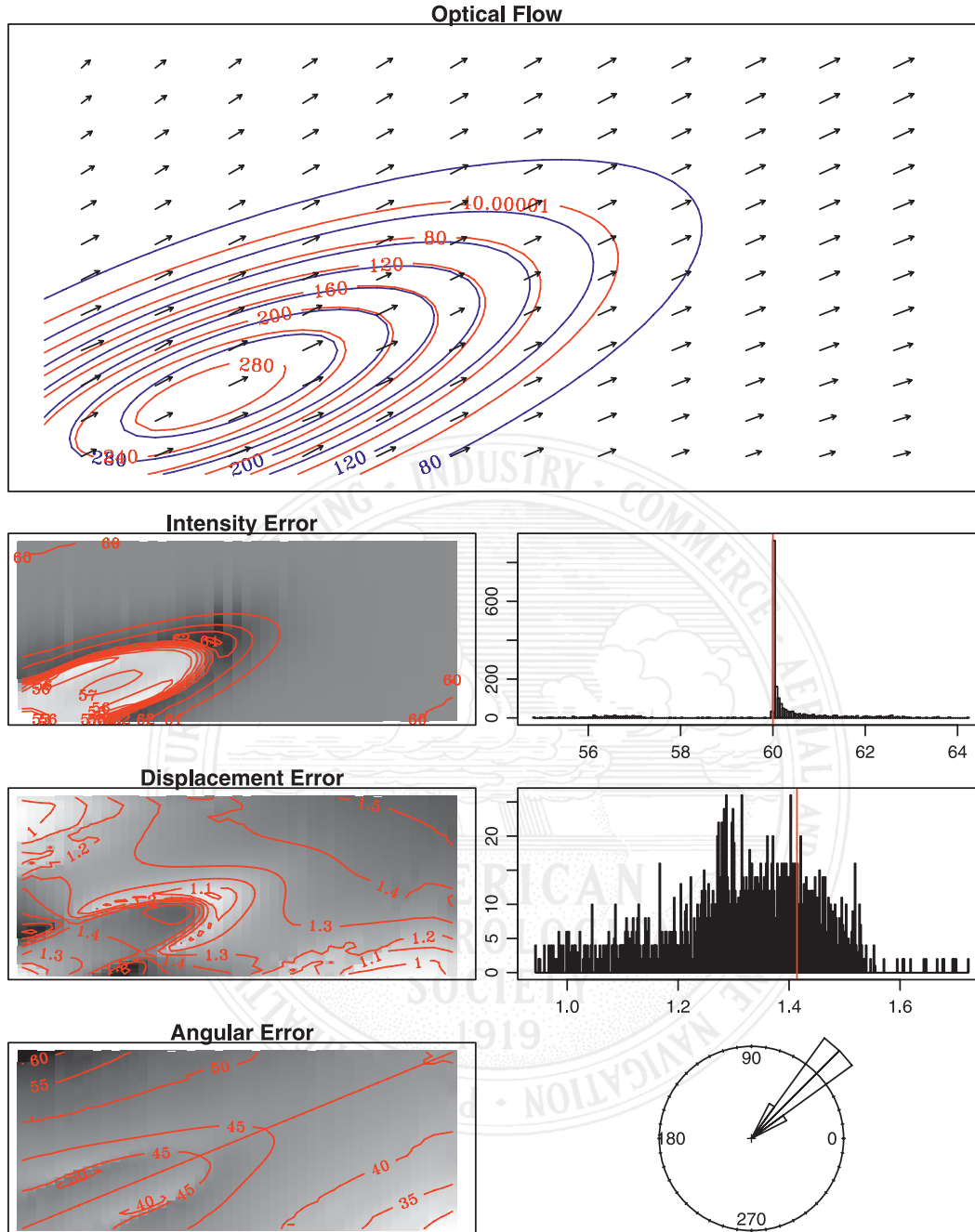


FIG. 1. The OF field and the three components of error in one of the simulated fields. (top) The OF field (arrows) mapping the forecast (red) to the observed (blue) field. (bottom) The (left) spatial distribution and (right) histogram of the three components of the forecast error. See text for details.

of the error is only 0.001 (i.e., 0.1%) for the nonlinear model. Regarding the magnitude of the displacement errors (PropDisErr), it can be seen that these errors are generally much larger than the intensity errors; in fact, for a shift of 10 grid lengths, the proportion of the error is nearly 5, or 500%, for the nonlinear model! As for angular errors, although PropAngErr is nearly 150% for a shift of

10 grid lengths, it is nearly 0% for shifts as large as two grid lengths. Note that although the nonlinear model performs better than the linear one in terms of intensity and displacement error, the two models are comparable in terms of estimating the angular errors.

Increasing the size of the window from 5 to 11 leads to similar results (three panels in the bottom/left in Fig. 2).

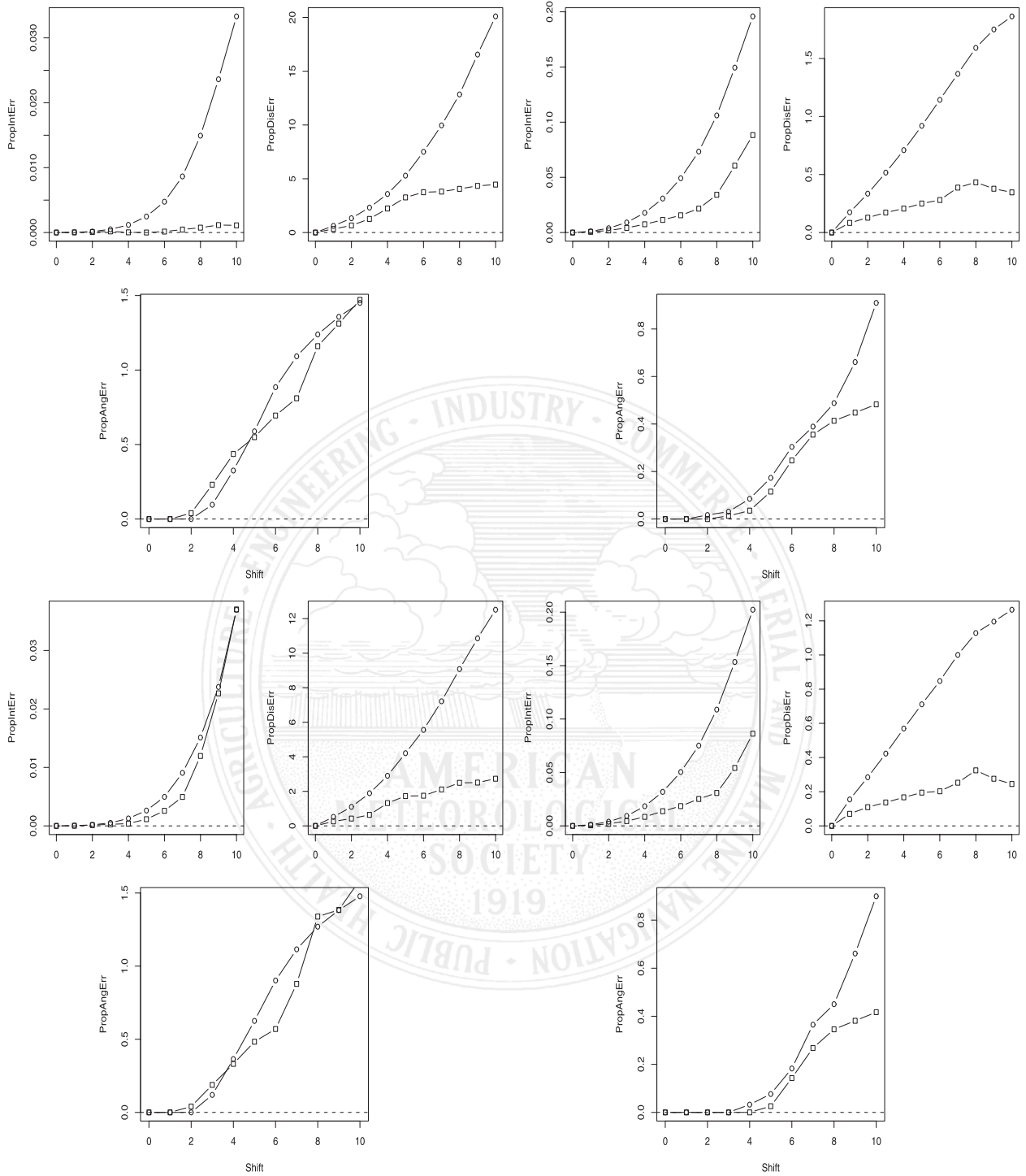


FIG. 2. The proportion of intensity error (PropIntErr), displacement (PropDisErr), and angular error (PropAngErr), as a function of the shift between the forecast and observed Gaussian fields, for two different Gaussian widths,  $S =$  (left) 7 and (right) 11, and at two different window sizes  $W =$  (top) 5 and (bottom) 11. Circles (squares) denote the linear (nonlinear) results.

One notable difference is that the nonlinear model does not add much to the linear model, in terms of estimating the intensity errors. As for the displacement errors, the nonlinear model is still superior. Also, the displacement errors are lower than that with  $W = 5$ . Based on an examination of this and other  $W$  values (not shown), it follows that for the parameters examined in this simulation, the choice of  $W$  does not significantly affect the results. In fact, as is shown next, the effect of  $W$  becomes even more muted as the size of the object increases.

For  $W = 5$  and  $S = 11$  (three panels in top/right of Fig. 2). The nonlinear model improves over the linear model in terms of all three measures. Here, PropIntErr is somewhat larger than that obtained for the smaller object with  $S = 7$ ; a shift of 10 grid lengths now leads to PropIntErr = 0.1 (i.e., 10%). However, PropDisErr and PropAngErr are both improved over their values for  $S = 7$ ; they are both at most about 0.4 (i.e., 40%). Finally, the  $W = 7, S = 11$  results (bottom/right in Fig. 2) are nearly identical to those of  $W = 5, S = 11$ . In other words, as mentioned above, the effect of  $W$  is reduced as the size of the feature increases.

As noted previously, one may estimate the true errors with the mode of the distribution rather than the median. The corresponding results are not shown here, but all of the estimates are improved. For example, over the range of shifts examined, the proportion of error never rises above 0.5 (i.e., 50%) for any component of error.

The nonlinear results appear a bit more “noisy” than the linear results. This is primarily the result of computational variance. An example of such a variance is that due to local minima. All nonlinear models have local minima of the error function, in which the system may get stuck. For realistic data, this variability due to local minima is not expected to be a source of concern, because it is usually overwhelmed by other sources of variation (e.g., sampling).

As discussed above, the joint histogram of the errors is a useful quantity. Figure 3 shows the joint histogram of the magnitude of displacement and angular errors for  $W = 9, S = 11$ , for a shift of one grid length in the  $x$  and  $y$  directions. These parameters are those pertaining to the simulation shown in Fig. 1. The distribution shows a clear peak occurring at the correct values of 1.41 and  $45^\circ$  for magnitude and angle, respectively. This type of correlation between angle and magnitude cannot be gathered from the marginal histograms. The joint histogram is, therefore, useful in assessing the “coherence” of the displacement and angular components of error.

The main purpose of this simulation study has been to assess what relative errors can be obtained, given the typical size of objects in a field and the window size  $W$ . However, it is important to point out that there is no

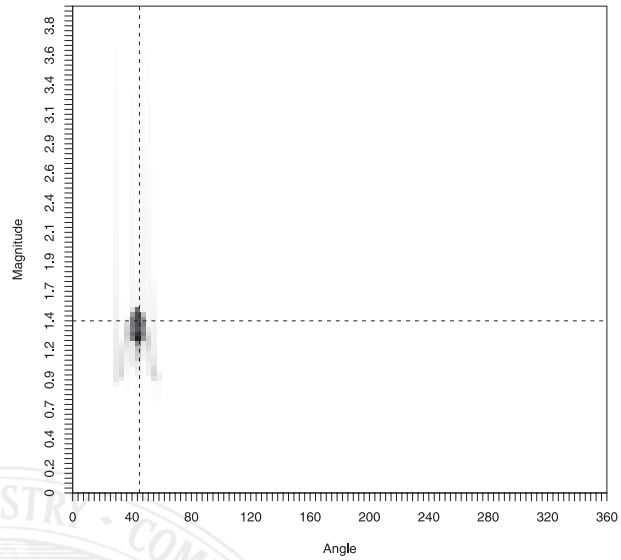


FIG. 3. The joint histogram of the displacement and angular errors corresponding to Fig. 1. The dashed lines indicate the true errors in magnitude and direction.

critical value of  $S$  (or  $W$ ) below which the OF results become unambiguously unacceptable; the results simply become more meaningful and useful as the field becomes smoother. In the next section, the OF method is applied to realistic data, and it is shown that meaningful results can indeed be obtained.

### 5. Results: Realistic data

In this section we apply the above-developed ideas to sea level pressure (SLP) data obtained from the University of Washington Mesoscale Ensemble (UWME) system; only one member is examined, namely the GFS–MM5 24-h forecasts. The domain is the 36-km domain, covered by a grid of  $116 \times 140$  grid points with a nominal 36-km grid spacing. First, a single forecast is examined for a 0000 UTC 4 December 2008 initialization. The UTC time corresponds to 1700 LST on the preceding day (i.e., 3 December). This date is selected because it corresponds to a major Pacific Northwest winter storm event during the 2007/08 winter. Figure 4 shows the forecast (top) and the corresponding observation/analysis (bottom). After an examination of this case, the method is applied to over a year’s worth of data (418 days) spanning the period from 2 April 2008 to 2 November 2009. Only the average of the OF field and the components of error are examined here.

Comparing the forecasts with the observations/analysis for 4 December carefully, one notes that, working west to east, the low pressure system on the Aleutians is forecast too far north and is too deep. The associated secondary

F3

F4



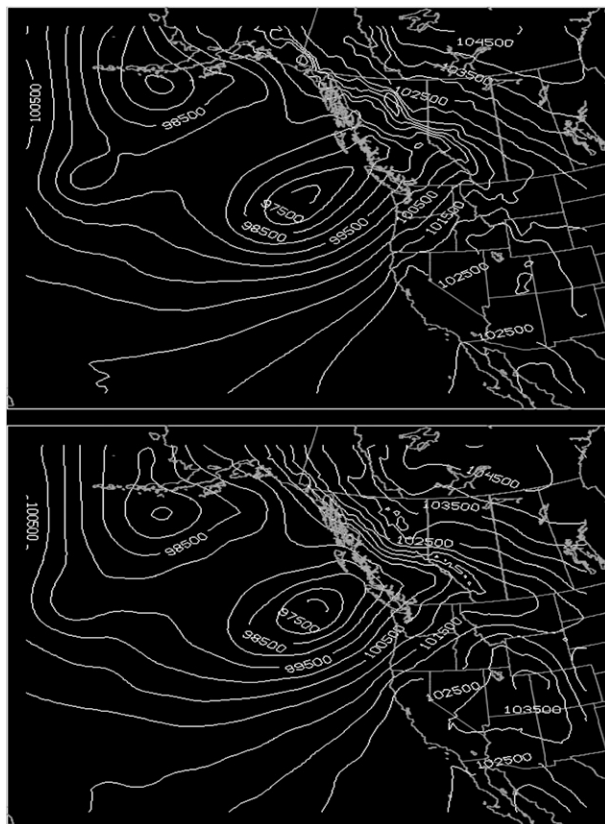


FIG. 4. The (top) forecast and (bottom) observations/analysis SLP fields for 0000 UTC 4 Dec 2008.

low to the southwest is also forecast too deep. The low pressure system off Vancouver, British Columbia, Canada, is too far southwest (or slow) and both this low and the associated troughing along the Canadian border appear to be forecast too deep while the ridge over the Canadian Rockies verifies much stronger than forecast. To the south, over the Utah Rockies, the high pressure center is also more intense than forecast by several hectopascals.

These errors can be seen readily in the error fields in Fig. 5. The window size  $W$  is set to 31, corresponding to a window of approximately 1000 km, which is representative of the synoptic features on the map. Smaller windows result in error fields that are more noisy, picking up many more small-scale features than are visible without very close examination. The error fields for larger values appear to be overly smoothed, indicating the features discussed but muting the errors by addressing too large an area. The next section presents more detailed OF results for different values of  $W$ .

Examining Fig. 5 more closely, in the histogram of intensity errors, and remembering what the values are (observed field – forecast field), the appearance of the mode and median to the right of zero (marked with

a red, vertical line) indicates that the forecasts are generally lower than the verifying value. Examining the spatial distribution of intensity errors, these are associated with underforecasting the high pressure areas over the Rockies and the strong gradient between the ridge and the coastal low, likely a model bias in this case. Interestingly, the entire ocean area, including the two systems mentioned, is slightly overforecast (higher pressure) than observed with the exception of the two low pressure systems; a conclusion that would bear careful examination if persistent over many cases. The histogram of the displacement errors implies that the typical displacement error is about one to two grid lengths. The largest displacement errors occur in the northwest corner of the region. According to the histogram of the angular errors, the OF vectors are directed in no single direction (i.e., there is no global shift error in the forecasts). There appear to be three (or four) modes in the histogram of the angles, corresponding approximately to the first, second, and third quadrants. Viewing the spatial pattern of the angles suggests that these three modes loosely correspond to three horizontal bands. Recalling that angles are computed counterclockwise from the east, it appears in this case that the model is exhibiting different biases in the subtropical, main westerlies and the more northern latitudes.

Figure 6 displays the average OF field and the three components of the error; the (vector) averaging is performed on the OF estimates from each of the 418 dates. Examining the OF diagram, it is evident that forecast errors are much more significant over land. The histogram of intensity errors indicates a model bias toward forecasting higher pressures than are observed. Examining the spatial distribution reveals little or no intensity bias in the upper latitudes, with a strong bias over southern California and also in the SW quadrant. The lack of a significant intensity bias in the upper latitudes may be due to the more transit nature of the synoptic features in the westerlies. The SW quadrant error, which also is present in the spatial displacement of the displacement error, could be a persistent boundary condition error or model issue with the placement of subtropical features. The spatial distribution also reveals a clear land–ocean difference, with spatial errors over the ocean (with the exception of the SW) generally insignificant but over land demonstrating a clear forecast bias. The angular error histogram indicates this bias is persistently to the northeast. The significant southern California errors in both intensity and location bear further examination and may indicate a serious model issue. The joint histogram of the averaged errors (Fig. 7) confirms that the above observations indicate that a significant number of one-grid-length displacements are associated with a

[F6]

[F7]

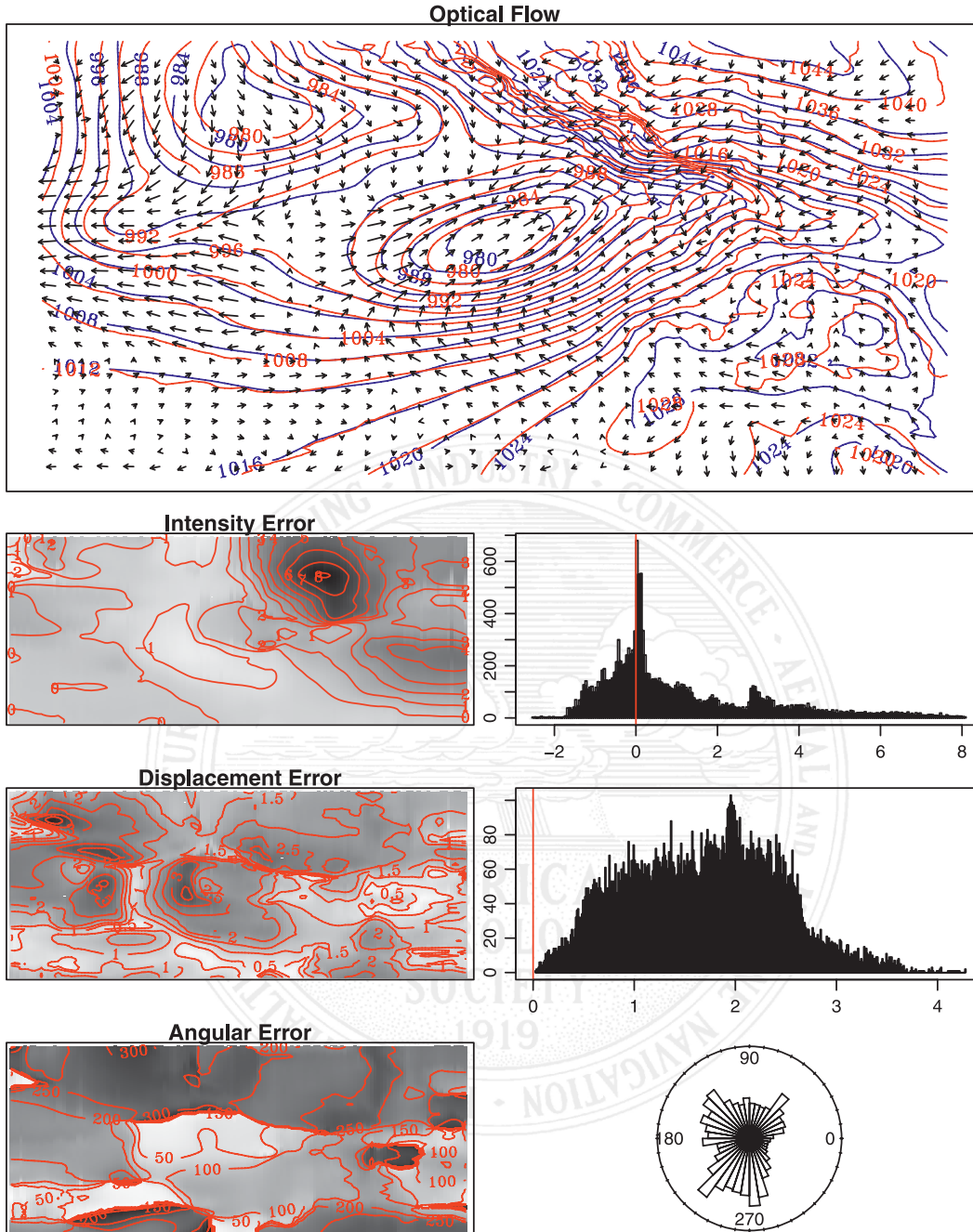


FIG. 5. As in Fig. 1, but for sea level pressure forecasts for 0000 UTC 4 Dec 2007, with  $W = 31$  (red = forecast; blue = analysis in the top panel.) See text for details. Sea level pressure is expressed in hPa.

unique and well-defined direction (west-southwesterly). Said differently, Figs. 6 and 7 suggest that the forecasts over land do have a general spatial bias in that they are placed about one grid length to the east-northeast of the observations. The forecast errors over water are much smaller in terms of their magnitude and more diffused in terms of their direction. As mentioned in the discussion

section, these smaller errors may not be statistically significant.

## 6. Choice of window size

As mentioned previously, this OF approach has a parameter  $W$  that must be specified. Recall that it is the

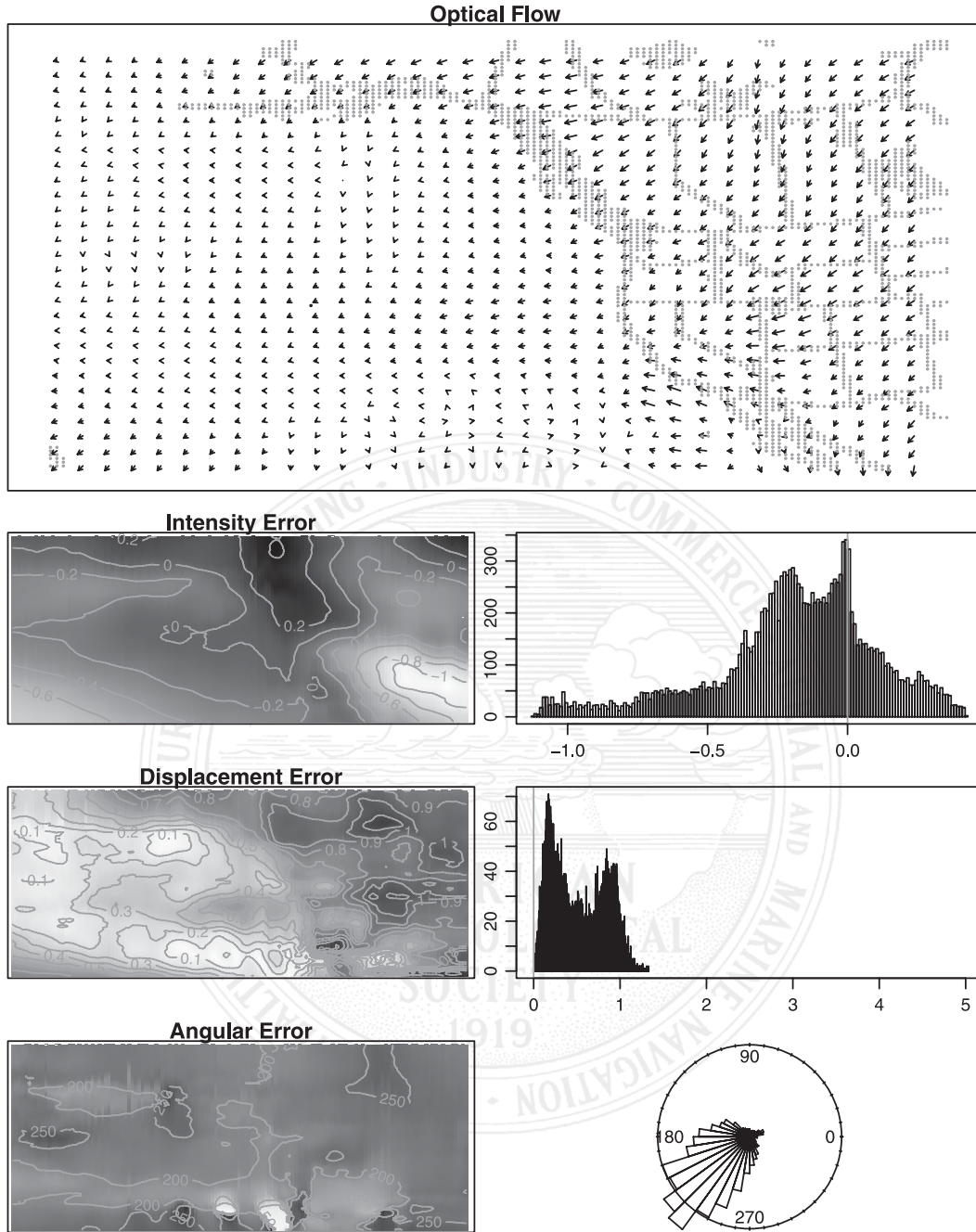


FIG. 6. The OF and the three error fields averaged over 418 days, at  $W = 31$ . In the contour plots, white (black) corresponds to low (high) values of the quantity being examined.

size of the window across which the OF field is assumed to be constant. As such, a sufficiently large window size is apt to lead to the violation of that assumption. On the other hand, a small window, with few pixels contributing to the estimation of the error parameters, can lead to poor estimates. In short,  $W$  is a smoothing parameter: increasing the size of the window has the effect of smoothing the OF field.

Before discussing how  $W$  should be selected, it helps if one has an intuitive sense of what  $W$  represents within a verification context. By choosing a specific value of  $W$ , one is effectively considering each image as being composed of an overlapping collage of square tiles, each with a constant intensity (i.e., no internal structure). The OF model, then, effectively slides these tiles, and changes their intensity, in an attempt to find the best match between the

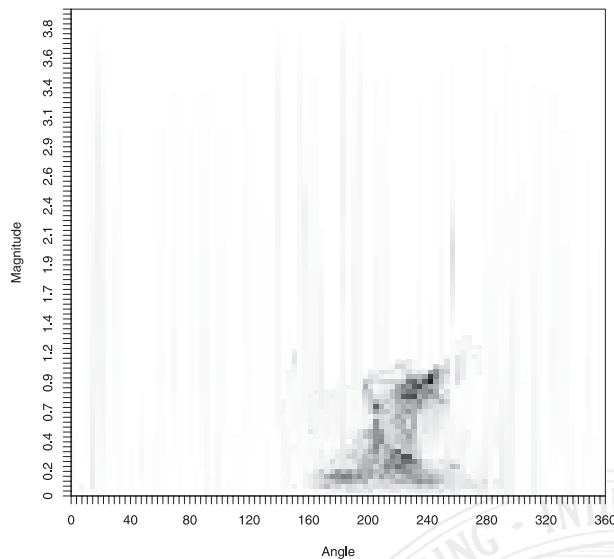


FIG. 7. The joint histogram of the displacement and angular errors averaged over 418 days, at  $W = 31$ .

forecast and the observed field. In this sense, the size of the window is a measure of the resolution/scale at which the verification is to be performed.

In practice, what value of  $W$  should one use? The answer depends on two scenarios. In one scenario, a user (e.g., operational forecaster) has a specific spatial scale of interest. In that case,  $W$  should be set to that value. In the second scenario, the user (e.g., a model developer or operational NWP center) has no specific scale in mind, but is interested in model performance and selection over a range of scales. In that case, the term “best” implies that the selected model is to be better at all scales, or at least over some range of scales. As such, the OF field must be examined for  $W$  values over that range. There may be other situations that call for a different treatment of  $W$ . Generally, however, checking the results for a range of  $W$  values makes for good practice, because at the least it gives the user an appreciation for how sensitive the errors are to the choice of scale. From a technical/computational point of view, this is not problematic because the estimation procedure can be set up such that for a given  $W$  the errors can be estimated for all smaller values of  $W$  sequentially, at no extra effort.

**F8** Figure 8 shows the distribution of the three components of error for 4 December 2007, at seven values of  $W$  shown along the right margin of the figure. The reason for the lowest value of  $W$  has been explained in section 3. The largest value ( $W = 101$ ) is a relatively arbitrary choice intended to examine the results if the window size is comparable to the size of the field itself ( $116 \times 140$ ). Also, for the sake of brevity, the spatial distributions are not shown.

The left column in Fig. 8 shows that the intensity errors are peaked at zero, implying that there is no intensity bias at any of the examined scales, as one would hope from NWP conservation assumptions. The spread of the histograms decreases with the size of the window. Said differently, at smaller scales, intensity errors may be larger. At  $W = 5$  (i.e., 180 km), intensity errors vary from approximately  $-5$  to 10 hPa, with a heavy tail out to 20 hPa. At the largest scale, the spread of the intensity errors is reduced to the range  $-1$  to 4 hPa. This pattern of behavior is expected, because at a specific value of  $W$ , all errors on scales smaller than  $W$  are smoothed out.

The magnitude of the displacement errors follows a similar pattern (Fig. 8, middle column). Over the full range of  $W$  values, the maximum of the errors goes from approximately six grid lengths (186 km) to about three grid lengths (98 km). The most likely error at the smallest scale is about two grid lengths, and interestingly, this error persists even at the largest scales. For  $W = 101$ , the appearance of a secondary mode in the histogram of displacement errors at about one grid length suggests that at this scale, the spatial field has two distinct regions, each with a different typical value of displacement error. This has been confirmed by examining the spatial distribution of the errors; the two modes at about one and two grid lengths correspond to the regions over water and land, respectively.

The distribution of the angular component of the displacement error is also dependent on  $W$ . On small scales more angles exist in the  $180$ – $270^\circ$  range. At the largest scale, the observed field appears to have one region to the southwest of the forecast field, and another region to the north of the forecast field. According to the spatial distribution of these errors (not shown), these two regions correspond to land and water, respectively.

It is important to point out that although only the extreme scales in Fig. 8 are examined here, the intermediate values of  $W$  are still useful—specifically, for the user interested in verification on the corresponding scales.

## 7. Statistical considerations

The formulation of the OF model as a parametric regression problem allows for “explanations” of some features in the results. For example, it suggests further revision of the model expressed in Eq. (5). Note that this regression equation is of the type  $y = \alpha + \beta_1 x_1 + \beta_2 x_2$ . It is easy to show that the least squares estimates of the regression coefficients  $\alpha, \beta_1, \beta_2$  are in general correlated. Specifically, the correlation between the estimates of  $\alpha$  and  $\beta_1$  is proportional to the sample mean of  $x_1$ . Similarly, the correlation between the estimates of  $\alpha$  and  $\beta_2$  is proportional to the sample mean of  $x_2$  (Draper and

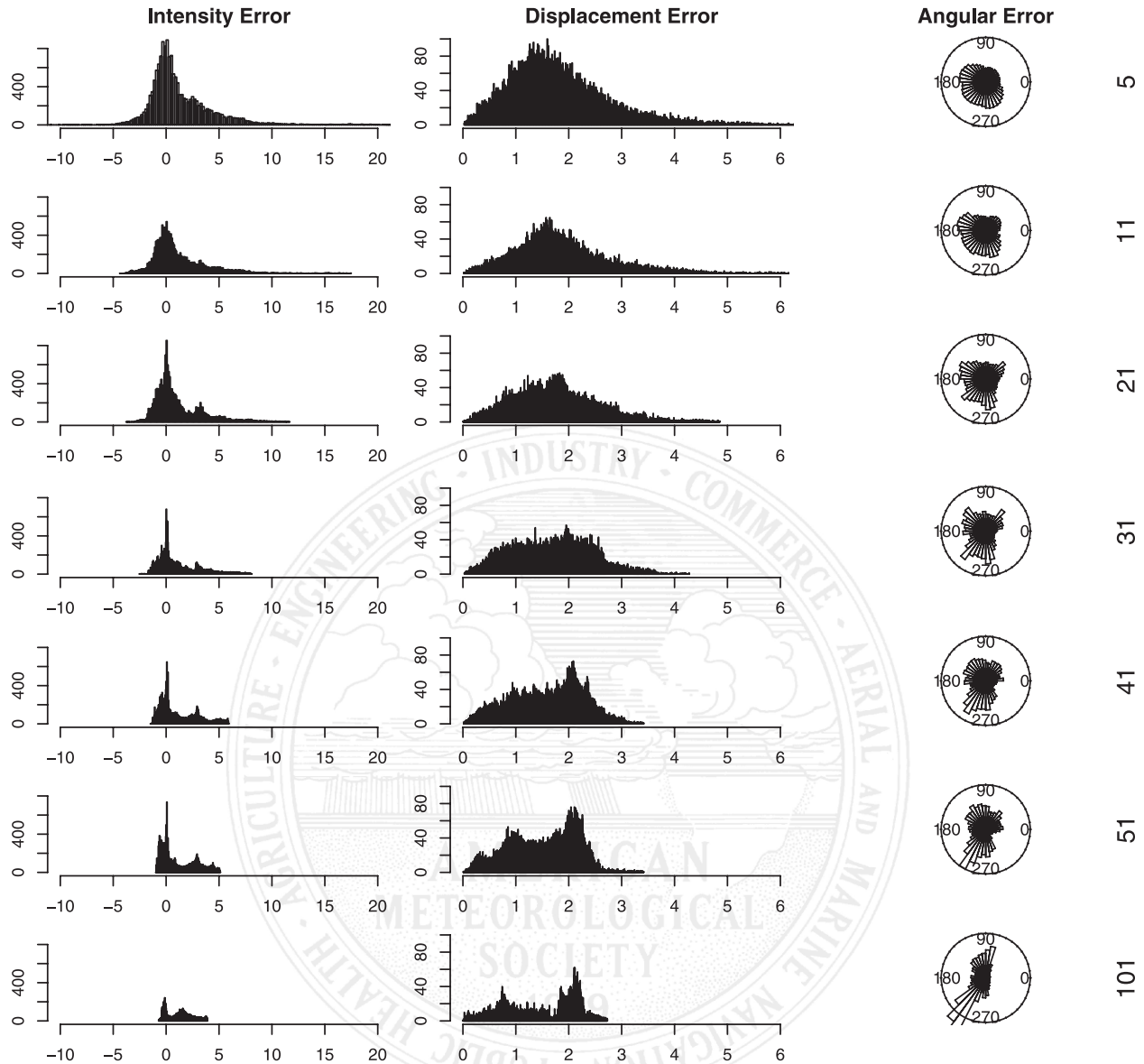


FIG. 8. The effects of the window size (denoted on the right of each row) on the distribution of the three components of the error. Data are SLP for 0000 UTC 4 Dec 2007.

Smith 1998, p. 129). This implies that the estimates of the intensity error will generally be correlated with the estimates of the displacement error in Eq. (5). Consequently, their interpretation becomes ambiguous in that one cannot be certain if the error is due to intensity or to displacement. This problem, however, is easily solved: in the current application, all of the predictors are centered by subtracting their respective means. The estimate of the intensity error is then assured to be uncorrelated with the estimates of the displacement error. Note that there may still exist an apparent correlation between these two quantities when one views their spatial distribution (e.g., Fig. 5), but that

correlation is a spatial one (i.e., a consequence of the underlying physics).

The regression formulation also aids in explaining some of the OF results. For instance, it explains why the OF field in Fig. 1 appears to be nonconstant across the field, even though the shift between the observed and the forecast object is a global one. Again, considering the regression equation  $y = \alpha + \beta_1 x_1 + \beta_2 x_2$ , it can be shown that the estimates of  $\beta_1$  and  $\beta_2$  are also correlated, with their correlation proportional to the sample correlation coefficient between  $x_1$  and  $x_2$ . This means that the direction and magnitude of the OF vectors cannot

be estimated unambiguously. An intuitive illustration of this ambiguity is found in the so-called aperture problem (Adelson and Movshon 1982). Consider an image consisting of diagonal, alternating black and white stripes. Now, allow the image to move in some direction. As seen through a small hole that reveals only a portion of the image, it is impossible to determine the direction in which the image has been moved. Note that if the stripes are nonlinear (e.g., curved), then it is possible to determine the direction of movement unambiguously. In other words, the ambiguity in the direction and magnitude of the OF vectors is present only in regions of the field where the forecast and observed objects have linear features. And this happens to be the case far away from the center of the Gaussians. This explains why the OF field in Fig. 1 appears distorted; far from the center of the Gaussian the field is mostly flat, leading to linear features that in turn render the estimation of  $dx$  and  $dy$  ambiguous. It is important to emphasize that the aperture problem is not a problem that can be solved by better models; as long as the data have linear features, then any model will suffer from this problem. Parametric models allow for an explicit explanation of the problem, while nonparametric models simply do not address it.

One might also wonder why the estimates of the intensity and displacement error in Fig. 1 are not exactly equal to the true values. Again, the regression interpretation of the OF model offers an explanation. First, note that the width (e.g., standard deviation) of the histogram reflects the precision of the estimate, while the difference between the center (e.g., median) of the histogram and the true value is the accuracy (or bias) of the estimate. In this language, while all three of the estimates are reasonably precise, the displacement and intensity errors are clearly biased. Given that the simulated forecast and observed field are completely consistent with the expression in Eq. (4), one would expect the estimate of the parameters to be unbiased. However, the model fitted to the data is not that in Eq. (4), but the approximate model appearing in Eq. (5). In other words, the postulated/fitted regression model is in fact not the correct/true model. It is known that this situation leads to biased estimates of the regression coefficients (Draper and Smith 1998, chapter 10). In other words, the bias (or inaccuracy) apparent in the simulated results (Fig. 1) is a direct consequence of neglecting higher derivatives. In realistic situations, one does not know the true model, and so this bias is usually ignored.

## 8. Summary and discussion

It is shown that a method designed for motion estimation, generally known as optical flow, can be developed to

assess the quality of forecasts of spatial fields, such as sea level pressure, in a diagnostic fashion. One traditional formulation, developed by Lucas and Kanade (1981), is extended to the nonlinear realm and is revised to allow for a decomposition of forecast error into three components: errors in intensity and two components of displacement—distance and angular error. It is also shown that the joint distribution of these errors conveys useful information. The method is illustrated on simulated data in order to examine the behavior of the results as a function of the typical size and separation of the spatial features, and their dependence on a parameter ( $W$ ) of the method that effectively fixes the spatial scale of interest. Then, operational forecasts of sea level pressure, averaged over 418 days, are examined within the proposed approach.

In the simulation study, since the truth is known, one can compute the accuracy of the estimated forecast errors. The accuracy of the error estimate is shown to improve with the accuracy of the forecast. There is no “critical value” of an object size or separation that marks the breakdown of the method; the accuracy of the estimated forecast errors simply diminishes with smaller objects, placed farther apart, as expected. These findings also depend on the parameter  $W$ , because it sets the scale over which the verification is to be performed.

The long-term analysis of sea level pressure forecasts demonstrates the value and ease of using this technique to verify and compare gridded forecasts. Results reveal a significant intensity bias in the subtropics, especially in the southern California region and in the southwest corner of the grid. It is not clear if the error in the southwest is due to the boundary conditions or to poor modeling of the subtropics. The analysis also exposes a systematic east-northeast or downstream bias of approximately 50 km over land, possibly due to the treatment of terrain in the coarse-resolution model. Finally, the joint distribution of intensity and displacement errors indicates that the displacement of the forecasts has a coherent spatial structure. In ongoing work the results of the OF analysis are being compared with human/expert assessment of the forecasts.

The intensity errors estimated in this paper are assumed to be purely additive. A model allowing for both additive and multiplicative errors was developed, specifically,  $I_o(x, y) \simeq A(x, y) + M(x, y) I_f(x + dx, y + dy)$ . It was found that the estimates of  $A(x, y)$  and  $M(x, y)$  are highly (and negatively) correlated. And unlike the additive model, centering the predictors does not eliminate the correlation. One possible explanation is that, over the examined scales, there is simply no way to distinguish between additive and multiplicative errors. To be able to distinguish between them, one would have to perform the analysis over a much larger scale (i.e., larger  $W$ ). But

that would tend to oversmooth the field. For this reason, in this work, only additive intensity errors are examined.

Several extensions of this work are worthwhile. For example, the error fields in Fig. 5 are all subject to sampling variability, and so it is important to associate some measure of statistical confidence to the estimate of the error at each grid point. Such an assessment of within-day variability will help in deciding, for example, whether or not a zero estimate truly is zero. A similar measure of confidence should also be computed for the average OF results shown in Fig. 6; that quantity will assess the between-day variability of the results and can help in comparing forecasts from different NWP models. Preliminary results based on the empirical sampling distribution of the errors indicate that some of the features in Fig. 6 may not be statistically significant. For example, errors in the magnitude of the displacement vectors around 0.2 hPa are consistent with the sampling distribution of such vectors mapping two random fields. Further work on the sampling distribution of the errors is currently in progress.

The simulation study performed here examines the dependence of the OF results on the typical size of an object; but the size is assumed to be the same in both the forecast and observed fields. In other words, it is assumed that objects do not change their size. A generalization of the proposed approach can introduce additional parameters that can represent size errors. Additionally, a useful study will be to assess the dependence of the results on the texture of the forecast and observed fields. In spatial statistics, texture is often quantified by a variogram (Marzban and Sandgathe 2009), which gauges the spatial extent of the correlations across the domain. It will be useful to set up a simulation where the OF error fields are examined as a function of texture.

Finally, here, the OF approach is applied to sea level pressure. The spatial continuity of the field simplifies the problem; mixed discrete-continuous fields, like reflectivity or precipitation, involve edges at which derivatives become ill-behaved. The simplest formulation of the Lucas-Kanade approach (i.e., linear and assuming no intensity error) typically gives rise to OF fields that are not readily interpretable (Marzban and Sandgathe 2007). Although this simple formulation can be applied to precipitation fields (Marzban et al. 2009), for such fields it is more appropriate to either employ nonparametric OF methods, such as in the method of Keil and Craig (2007), or to preprocess the forecast and observed fields in order to smooth the edges of the objects in the domain. The nonparametric approach has neither the benefits nor the transparency of the parametric approach developed here, but it will be interesting to compare the approaches in terms of their performance. The application of the

proposed approach to precipitation fields is currently under investigation.

*Acknowledgments.* V. Lakshmanan, Nicholas Lederer, Dustin Lennon, Werner Stuetzle, Don Percival, and Albert Kim are acknowledged for contributing to this work. Partial support for this project was provided by National Science Foundation Grant 0513871 and Office of Naval Research Grants N00014-01-G-0460/0049 and N00014-05-1-0843.

## REFERENCES

- Adelson, E. H., and J. A. Movshon, 1982: Phenomenal coherence of moving visual patterns. *Nature*, **300**, 523–525.
- Baker, S., and I. Matthews, 2004: Lucas-Kanade 20 years on: A unifying framework. *Int. J. Comput. Vis.*, **56**, 221–255.
- Bowler, N. E. H., C. E. Pierce, and A. Seed, 2004: Development of a precipitation nowcasting algorithm based on optical flow techniques. *J. Hydrol.*, **288**, 74–91.
- Brewster, K. A., 2003: Phase correcting data assimilation and application to storm-scale numerical prediction. *Mon. Wea. Rev.*, **131**, 480–507.
- Brill, K. F., 2002: Automated east-west phase error calculation. HPC Tech. Rep., NOAA/NSW/NCEP/HPC. [Available online at [http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/Brill/Brill\\_EW\\_PhaseError.html](http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/Brill/Brill_EW_PhaseError.html).] **AU1**
- Brown, B. G., and Coauthors, 2004: New verification approaches for convective weather forecasts. Preprints, *11th Conf. on Aviation, Range, and Aerospace*, Hyannis, MA, Amer. Meteor. Soc., 9.4. [Available online at <http://ams.confex.com/ams/pdfpapers/82068.pdf>.]
- Casati, B., and Coauthors, 2008: Forecast verification: Current status and future directions. *Meteor. Appl.*, **15**, 3–18.
- Chan, E., 1993: Review of block matching based motion estimation algorithms for video compression. *Canadian Conf. on Electrical and Computer Engineering*, Vancouver, BC, Canada, 151–154. **AU2**
- Davis, C., B. Brown, and R. Bullock, 2006: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev.*, **134**, 1772–1784.
- Douglas, R. J., 2002: Rearrangements of functions with applications to meteorology and ideal fluid flow. *Large-Scale Atmosphere-Ocean Dynamics: Analytical Methods and Numerical Models*, J. Norbury and I. Roulstone, Eds., Cambridge University Press. **AU3**
- Draper, N. R., and H. Smith, 1998: *Applied Regression Analysis*. 3rd ed. John Wiley and Sons, 706 pp.
- Du, J., S. L. Mullen, and F. Sanders, 2000: Removal of distortion error from an ensemble forecast. *Mon. Wea. Rev.*, **128**, 3347–3351.
- Gebremichael, M., W. F. Krajewski, and G. Ciach, 2004: Assessment of the statistical characterization of small-scale rainfall variability from radar: Analysis of TRMM ground validation datasets. *Sixth Int. Symp. on Hydrological Applications of Weather Radar*, Melbourne, VIC, Australia. **AU4**
- Germann, U., and J. Joss, 2001: Variograms of radar reflectivity to describe the spatial continuity of Alpine precipitation. *J. Appl. Meteor.*, **40**, 1042–1059.

- , and I. Zawadzki, 2002: Scale-dependence of the predictability of precipitation from continental radar images. Part I: Description of the methodology. *Mon. Wea. Rev.*, **130**, 2859–2873.
- , and —, 2004: Scale-dependence of the predictability of precipitation from continental radar images. Part II: Probability forecasts. *J. Appl. Meteor.*, **43**, 74–89.
- Hodges, K. I., 1994: A general method for tracking analysis and its application to meteorological data. *Mon. Wea. Rev.*, **122**, 2573–2586.
- , 1995: Feature tracking on the unit sphere. *Mon. Wea. Rev.*, **123**, 3458–3465.
- , 1996: Spherical nonparametric estimators applied to the UGAMP model integration for AMIP. *Mon. Wea. Rev.*, **124**, 2898–2913.
- , 1998: Feature-point detection using distance transforms: Application to tracking tropical convective complexes. *Mon. Wea. Rev.*, **126**, 785–795.
- , 1999: Adaptive constraints for feature tracking. *Mon. Wea. Rev.*, **127**, 1362–1373.
- Hoffman, R. N., and C. Grassotti, 1996: A technique for assimilating SSM/I observations of marine atmospheric storms: Tests with ECMWF analyses. *J. Appl. Meteor.*, **35**, 1177–1188.
- , Z. Liu, J.-F. Louis, and C. Grassotti, 1995: Distortion representation of forecast errors. *Mon. Wea. Rev.*, **123**, 2758–2770.
- Horn, B. K. P., and B. G. Schunck, 1981: Determining optical flow. *Artificial Intell.*, **17**, 185–203.
- 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.
- , and —, 2009: A displacement and amplitude score employing an optical flow technique. *Wea. Forecasting*, **24**, 1297–1308.
- Lim, D.-K., and Y.-S. Ho, 1998: A fast block matching motion estimation algorithm based on statistical properties of object displacement. *Proc. 2001 Int. Symp. on Intelligent Multimedia, Video and Speech Processing*, Hong Kong, China, IEEE, 498–501.
- Lucas, B. D., and T. Kanade, 1981: An iterative image registration technique with an application to stereo vision. *Proc. Imaging Understanding Workshop*, 121–130. **AU5**
- Marzban, C., and S. Sandgathe, 2007: Verification with optical flow. *Proc. Third Int. Workshop on Verification Methods*, Reading, United Kingdom, ECMWF. [Available online at <http://ecmwf.int/newsevents/meetings/workshops/2007/jwgv/index.html>.] **AU6**
- , and —, 2009: Verification with variograms. *Wea. Forecasting*, **24**, 1102–1120.
- , —, H. Lyons, and N. Lederer, 2009: Three spatial verification techniques: Cluster analysis, variogram, and optical flow. *Wea. Forecasting*, **24**, 1457–1471.
- Nehrkorn, T., R. N. Hoffman, C. Grassotti, and J.-F. Louis, 2003: Feature calibration and alignment to represent forecast errors: Empirical regularization. *Quart. J. Roy. Meteor. Soc.*, **129**, 195–218.
- Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, 1999: *Numerical Recipes in C: The Art of Scientific Computing*. 2nd ed. Cambridge University Press, 994 pp.
- R Development Core Team, cited 2008: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. [Available online at <http://www.R-project.org/>.]
- Reilly, C., P. Price, A. Gelman, and S. Sandgathe, 2003: Using image and curve registration for measuring the goodness of fit of spatial and temporal predictions. *Biometrics*, **60**, 954–964. **AU7**
- Trucco, E., and A. Verri, 1998: *Introductory Techniques for 3-D Computer Vision*. Prentice Hall, 343 pp.

