Analyzing the Image Warp Forecast Verification Method on Precipitation Fields from the ICP

ERIC GILLELAND
Research Applications Laboratory, National Center for Atmospheric Research,* Boulder, Colorado

JOHAN LINDSTRÖM
Mathematical Statistics, Centre for Mathematical Sciences, Lund University, Lund, Sweden, and Department of Statistics, University of Washington, Seattle, Washington

FINN LINDGREN
Mathematical Statistics, Centre for Mathematical Sciences, Lund University, Lund, Sweden

(Manuscript received 15 October 2009, in final form 12 February 2010)

ABSTRACT

Image warping for spatial forecast verification is applied to the test cases employed by the Spatial Forecast Verification Intercomparison Project (ICP), which includes both real and contrived cases. A larger set of cases is also used to investigate aggregating results for summarizing forecast performance over a long record of forecasts. The technique handles the geometric and perturbed cases with nearly exact precision, as would be expected. A statistic, dubbed here the IWS for image warp statistic, is proposed for ranking multiple forecasts and tested on the perturbed cases. IWS rankings for perturbed and real test cases are found to be sensible and physically interpretable. A powerful result of this study is that the image warp can be employed using a relatively sparse, preset regular grid without having to first identify features.

1. Introduction

As forecast models have progressed to higher-resolution grids, their usefulness to most users has increased. Unfortunately, traditional forecast verification scores [e.g., root-mean-square error (RMSE), probability of detection (POD), etc.; Jolliffe and Stephenson (2003); Wilks (2006)] that are calculated on a gridpoint-to-gridpoint basis often conclude that the models do not perform as well as lower-resolution models. For example, increased small-scale variability results in greater occurrences of small errors. There is also the double-penalty issue whereby a small spatial displacement contributes negatively for both false alarms and misses (i.e., for the same error). Additionally, the traditional verification scores provide only limited diagnostic information about how the forecast performed (e.g., What did the forecast get right or wrong? Are there geographic regions where the forecast performs better or worse?). These issues have prompted numerous new verification methods to be proposed. Many of the methods are summarized in Gilleland et al. (2009), and they fall roughly into four categories: (i) features based (e.g., Ebert and McBride 2000; Ebert and Gallus 2009; Davis et al. 2006, 2009; Micheas et al. 2007; Wernli et al. 2008, 2009; Venugopal et al. 2005), (ii) scale decomposition (e.g., Briggs and Levine 1997; Zepeda-Arce et al. 2000; Harris et al. 2001; Casati et al. 2004; Casati 2010; Lack et al. 2010), (iii) neighborhood (e.g., Ebert 2008, 2009; Mittermaier and Roberts 2010, and the references therein), and (iv) field deformation (e.g., Alexander et al. 1999; Hoffman et al. 1995; Keil and Craig 2007, 2009; Marzban et al. 2009; Nehrkorn et al. 2003; Reilly et al. 2004; Sampson and Guttorp 1999). Of course, some methods proposed do not fall as nicely into these categories.
categories (e.g., Brill and Mesinger 2009; Lakshmanan and Kaín 2010; Marzban and Sandgate 2006, 2008, 2009; Mesinger 2008; Nachamkin 2004, 2009).

In Åberg et al. (2005) an image-warping approach to short-term precipitation forecasts was introduced. Here, a similar method is applied for forecast verification. The warping deforms a set of forecast fields to match a verification field more closely. This gives an estimate of the spatial errors, represented by the warping deformations of the forecasts (i.e., errors in location, orientation, coverage area, and areal shape), as well as the reduction in intensity. The method can also be used to verify forecasts and is similar in some respects to the procedure of Keil and Craig (2007, 2009), as well as the optical flow method in Marzban et al. (2009), but with some important differences. In particular, the deformed forecast is obtained through a warping function as opposed to the hierarchical movement of points. Further, because the image-warping approach follows a stochastic model, there is a natural formulation for calculating uncertainty information for parameters of interest (e.g., spatial and intensity errors). The image-warping procedure described here is more closely related to the methods introduced independently by Alexander et al. (1999), Hoffman et al. (1995), Nehrkorn et al. (2003), Reilly et al. (2004), and Sampson and Guttorm (1999); the primary differences being the method for choosing the subset of points for fitting the image warp function (or control points), and specific choices of warping functions and their associated likelihoods (see sections 2a and 2b).

Although the procedure is automatic in that once it is begun it does not require any further user input, there are user-defined choices that can affect the resulting deformations. Specifically, one must (i) choose how many control points to use and where to place them, (ii) construct an appropriate loss function, and (iii) decide how to penalize components of the loss function in order to obtain useful information about forecast performance. The objective of this paper is to analyze the sensitivity of the method to the number of control points and the choice of prior information on the loss function (e.g., high versus low penalty on nonlinear deformations). We also demonstrate how to use the procedure, and how to interpret the results for various forecasting examples. The answers to these questions may depend on specific forecast user needs, so an attempt is made here to instruct any user on the decision process, and which issues are most important for gleaning useful information. We focus on the forecasting of precipitation fields because this has been the primary focus for the majority of the spatial forecast verification techniques proposed so far. However, the procedure works on a wide variety of field types (e.g., wind vectors, aerosol optical thickness, and binary fields). Hoffman et al. (1995) briefly discuss several types of displacement methods including optical flow and image warping, referred to therein as representation and variational analysis, respectively. However, the examples they give utilize a different displacement method. It should be noted that the displacement methods, including this one, characterize spatial location errors, and some summary for intensity error (e.g., RMSE). However, the methods allow for more informative information about intensity errors, such as the reduction in intensity error after applying the field displacement used here.

The next section gives a brief review of the image warping method used here. More technical details can be found in J. Lindström et al. (2010, unpublished manuscript, henceforth referred to as LGL; available online at http://www.ral.ucar.edu/staff/ericg/LindstromEtAl2009.pdf). Section 3 explains the details of the particular forecast and verification fields employed in this paper. Results are then described in section 4, followed by a summary, conclusions, and some discussion of issues and future work in section 5.

2. Image-warping verification

Image warping is briefly described here to give an understanding of the method, and the reader is referred to LGL for more detailed technical information. The method is based on a stochastic model consisting of three parts: (i) a warping function $W$ that controls the deformation of the forecast field; (ii) an error function $\varepsilon$ that models the intensity deviation between the deformed forecast and the verification field; and (iii) a smoothness prior on the deformations $g(W)$ that penalizes unrealistic deformations. Subsequently, as will be seen in this section, the warp function utilizes a conditional likelihood of the verification field given the forecast field that is consistent with the general framework for forecast verification proposed by Murphy and Winkler (1987).

The stochastic model is described in section 2a, and an estimation procedure for the model parameters is described in section 2b followed by specific considerations for obtaining useful information about forecast performance in section 2c. Finally, a ranking algorithm is proposed in section 2d.

a. Stochastic model

We denote the (undeformed) forecast and verification fields as $F$ and $O$, respectively. The stochastic model is then given by

$$O(s) = F[W(s)] + \varepsilon(s), \text{ for all } s \in D, \quad (1)$$
where $D$ is the entire domain of the verification and (undeformed) forecast fields; $W$ takes coordinates, $s = (x, y)$, from the deformed image field, denoted here by $\tilde{F}(s) = F(W(s))$, and maps them to coordinates in the undeformed field; and $\epsilon(s)$ are random errors.

The model (1) states that the verification field is a deformation of the forecast field plus some added innovations, representing amplitude errors that remain after the deformation.

The interpretation of the image warping is that it maps each point in one image to a point in another image. A pixel at point $s$ in the new image $\tilde{F}$ is assigned the pixel value of the undeformed forecast $F$ taken at $W(s)$. Although $W$ defines a mapping of any point $s$ in the continuous domain of the image, the deformation is completely defined by how a finite subset of points, $p_1 = (x_1, y_1), \ldots, p_n = (x_n, y_n)$, are mapped. Here, we refer to these points as control points, but in other contexts they are sometimes referred to as landmarks or tie points. Fixing a set of $n$ control points $p_{\text{obs}}$ in $O$, we require that the warping function fulfills

$$[p_1^{\text{fcst}}, \ldots, p_n^{\text{fcst}}] = [W(p_1^{\text{obs}}), \ldots, W(p_n^{\text{obs}})].$$

It should be observed that this is a requirement on the warping function and not a definition of the points. Here, we will use the common choice of a thin-plate spline for the warping function $W$ (e.g., Sampson and Guttorp 1992; Glasbey and Mardia 2001; Åberg et al. 2005), although several other families of transformations exist (for some alternatives, see Goshtasby 1987; Lee et al. 1997; Glasbey and Mardia 1998). Further, we fix the points in the verification field to lie on a regular grid (we feel that this makes more sense than fixing the points in the forecast field because we typically want to compare several forecasts to one verification field). Given the requirement (2), finding an optimal deformation is equivalent to finding points in the forecast field, which minimize the likelihood function defined below.

Two issues that arise in practice are that (i) the image fields are represented by pixels on a discrete grid and $W(s)$ may not coincide with a discrete pixel exactly and (ii) it is possible for a point to be mapped outside of the domain of the image. In the first case, it is sufficient to interpolate the points. Åberg et al. (2005) employ a bilinear interpolation, but we choose to use a cubic interpolation in order to facilitate a more efficient optimization routine. For the second issue, it is customary to either set the values to zero (cf. Glasbey and Mardia 2001) or extrapolate the values using the best linear unbiased estimator [i.e., kriging; cf. Åberg et al. (2005)]. For the precipitation fields tested here, it is appropriate to use zeros outside the image domain because such fields tend to have a lot of zero values.

Having fixed the control points $p_{\text{obs}}$ we arrive at a conditional distribution $g$ for $O$ describing the error likelihood. That is, the distribution for $\epsilon(s) = O(s) - F(W(s))$ in (1). Namely,

$$g(O|F, p_{\text{obs}}, p^{\text{fcst}}, \theta) = g(\epsilon|\theta),$$

where $\theta$ are parameters of the likelihood that need to be estimated. That is, the conditional distribution for $O$ given the forecast field, locations of the control points, and parameters of the error distribution $\epsilon$ conditional on $\theta$. Note finally that the error likelihood (3) is consistent with the general framework for forecast verification proposed by Murphy and Winkler (1987).

Selection of the specific error distribution depends on the type of field being studied. For example, Åberg et al. (2005) used Gaussian errors. For the quantitative prediction forecasting (QPF) fields, a mixture of two Gaussians on the difference between deformed forecast and observations gives a more accurate description of the errors, largely because of the many contiguous grid points where the values are zero. To better ensure the validity of the Gaussian assumption for the error field, a cubed root transformation is first applied to $O$ and $F$ (cf. Glasbey and Nevison 1997). This transformation is purely for statistical purposes and does not affect the physical interpretations of the resulting warp. Consequently, the error distribution used here is

$$\epsilon(s) \sim (1 - \pi)N(0, \sigma^2) + \pi N(0, \sigma^2 5 \times 10^{-5}),$$

where the mixture probability $\pi$, and variance $\sigma^2$, make up the parameters $\theta$ to be estimated. These parameters could be thought of as roughly the proportion of non-zero grid points and the variance of the residual field, respectively. However, their importance is in obtaining a correct deformation, and we consider them to be nuisance parameters. The above equation attempts to model the error distribution as a mixture of zeros (because of numerous zeros in both the verification and forecast field) and actual intensity errors (i.e., amplitude errors, misses, false alarms, etc.). The second component of the mixture has a much smaller variance than the first component in order to capture the zeros. The value of $5 \times 10^{-5}$ was chosen to allow the component to model the zero errors without causing numerical problems in the estimation.

To control against nonphysical, nonsmooth, or uninterpretable deformations, we impose a prior model on the warping function $W$. That is, we want to ensure that the deformation behaves consistently with how a human
observer would be likely to mentally morph the forecast image. For example, it is important to obtain a warp resistant to folding (i.e., deformations resulting in grid points’ being swapped with each other) because this type of behavior does not provide useful information. Further, depending on the user or type of forecast, one may want to put more or less emphasis on certain deformations. For example, one might not want the warp to move a forecast mass further than a few kilometers if it is reasonable to assume that such a discrepancy would really not be a spatial displacement error but simply, for example, an overforecast or false alarm type of error. Of course, the penalty only makes such deformations less likely to occur, and they can still result even with a high penalty. For brevity, only a penalty on the total amount of deformation (or bending energy) is applied [see Sampson and Guttorp (1992), Glasbey and Mardia (2001), and Åberg et al. (2005) for other possible penalties.]

Because the behavior of the warping function is determined by the positions of the control points, \( p^{\text{obs}} \) and \( p^{\text{fcst}} \), the prior model is simply a distribution on these points. Additionally, the control points for the observation field are considered to be fixed and known a priori so that the prior on the warping function is completely described by

\[
g(p^{\text{fcst}}|p^{\text{obs}}). \tag{4}
\]

Because we are using a penalty based on the bending energy, it can be shown that the appropriate distribution of the error is

\[
p^{\text{fcst}}|p^{\text{obs}} \in N(p^{\text{obs}}, \frac{1}{\beta} L),
\]

where \( \beta \) is the size of the penalty (large values correspond to small deformations) and \( L \) is the bending energy matrix of the thin-plate splines, which can be found in, for example, Dryden and Mardia (1998).

b. Estimation

The unknowns that must be estimated include the points \( p^{\text{fcst}} \) (i.e., the points warped from \( O \)) and the parameters \( \theta \). The knowns are considered to be \( O, F, \) and the control points \( p^{\text{obs}} \). Therefore, we seek the likelihood \( g(p^{\text{fcst}}, \theta|O, F, p^{\text{obs}}) \). Using the definition of conditional probability (i.e., \( \Pr(A, B) = \Pr(A|B)\Pr(B) \)) and Bayes’s rule (i.e., \( \Pr(A|B) \propto \Pr(B|A)\Pr(A) \)), and assuming that \( p^{\text{fcst}} \) is independent of \( F, \) conditional on \( p^{\text{obs}} \), the distributions (3) and (4) lead to the posterior

\[
g(p^{\text{fcst}}, \theta|O, F, p^{\text{obs}}) \propto g(O|F, p^{\text{fcst}}, p^{\text{obs}}, \theta)g(p^{\text{fcst}}|p^{\text{obs}})g(\theta)
\]

\[
= g(\varepsilon(\theta))g(p^{\text{fcst}}|p^{\text{obs}})g(\theta). \tag{5}
\]

Estimation is carried out by optimizing (5) with respect to \( p^{\text{fcst}} \) and \( \theta \). Note the additional term for the prior on \( \theta \). Because we do not have any prior information for these parameters, we use a flat prior so that \( g(\theta) = 1 \). The exact mathematical details of this are beyond the scope of this paper. However, the calculations of the likelihood and derivatives are very similar to those in Åberg et al. (2005) although we use a slightly different error distribution. Exact details can be found in LGL.

A broad outline of the optimization is as follows:

(i) Control points, \( p^{\text{obs}} \), are placed on a regular grid in the verification field.

(ii) The fields, \( F \) and \( O \), are smoothed by convolution with a Gaussian kernel.

(iii) The log-likelihood [i.e., the logarithm of (5)] and analytical first and second derivatives of the log-likelihood are calculated.

(iv) Given analytical derivatives, a Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm (algorithms 2.6.2 and 2.6.4 in Fletcher 1987) is used to find the maximum. The optimization essentially moves the control points \( p^{\text{fcst}} \) (and alters the parameters \( \theta \)) until a deformation that reduces the error between forecast, \( F \), and verification field, \( O \), while at the same time not incurring a too large deformation penalty \( g(p^{\text{fcst}}|p^{\text{obs}}) \) has been found.

(v) Steps ii–iv are repeated using successively smaller convolution kernels, and each time starting the optimization from the maximum found in the previous step iv.

Matlab routines are employed for the analysis here, and will be made available at the Spatial Forecast Verification Intercomparison Project (ICP) Web site (http://www.ral.ucar.edu/projects/icp/).

c. Forecast verification considerations

To capture as much information about forecast performance as possible, it is useful to calculate the base intensity error, \( \varepsilon_0 \), between \( F \) and \( O \) before deforming the forecast field, which is equivalent to the traditional verification procedure. We use the notation

\[
\varepsilon_0 = \varepsilon_0(F, O)
\]

to emphasize on which two fields \( \varepsilon_0 \) is applied, and the boldface type emphasizes that a vector of intensity errors could be used, but here we are concerned only with the RMSE.

After having applied the warping function, the reduction in the base intensity error is computed by calculating the intensity error, \( \varepsilon_1 = \varepsilon_1(F|W(s), O) \). A more useful measure of the improvement in forecast error is
the error reduction, \((e_0 - e_1)/e_0 \times 100\%\), rather than \(e_1\) itself. Loosely speaking, the error reduction can be interpreted as the percentage of the original error that is explained by the deformation. Being a percentage, the error reduction will be invariant to the size of the original intensity error.

Different types of deformations are utilized with image warping, and can be divided into two categories as affine and nonlinear. The affine deformations capture larger-scale errors and give information about spatial displacement, rotations, and global scaling errors. The nonlinear deformations provide localized error information, but are not as easy to interpret. It is possible to extract information about these two types of deformations separately, and might be of practical use for certain users. However, because such results will be sensitive to the choice of prior penalty placed on the nonlinear (or even certain affine) transformations, we do not feel that it is proper to explore this issue here.

d. Relative performance algorithm

It is often desired to compare and rank the performance of two or more forecast models. In the univariate domain, objective performance measures, such as the Gilbert skill score (GSS) and bias, are compared by taking the differences in these measures against zero using confidence intervals (or hypothesis testing) to determine the statistical significance of the result. The question is not always clear in the spatial domain and largely depends on user-specific requirements.

The image warp provides information about location errors, and when used in conjunction with an intensity error statistic leads also to information about how much better the forecast could have been. In the case of a good forecast with a small timing error, for example, the amount of displacement will be small and the reduction in error will be high. Such information is helpful in diagnosing a single forecast’s performance at a single point in time. However, it is not straightforward to summarize these scores to derive a ranking of performance for multiple forecasts. We therefore propose an algorithm for ranking one or more forecasts, which we will call here the image warp statistic (IWS). Define

\[
IWS_j = c_{1j}D_j + c_{2j}(1 - \eta_j) + c_{3j}\text{AMP}_j, \tag{6}
\]

where \(j\) indexes the forecasts being compared, \(D_j\) is the (standardized) average movement of points in both the \(x\) and \(y\) directions resulting from the image warp, \(\eta_j\) measures the amount of reduction in error, AMP\(_j\) is a measure of the original intensity error (here, the standardized original RMSE is used), and \(c_{ij} - i = 1-3, \ j \) indexes the forecast—are user-chosen weights that should vary depending on the values of \(D_j\) and \(\eta_j\). General guidance of how these weights might be chosen is given here, and section 4b gives an example to further illustrate the idea.

Selection of the weights, \(c_{ij}\), is complicated by the issue of whether a forecast error is a location (e.g., timing) error, or whether it is a miss or false alarm. The answer to such a question is likely to be subjective, but the assumption here, and generally for field displacement methods, is that small displacements of the forecasts are location (or spatial extent) errors while large displacements may be a miss and false alarm. This is accomplished with the image warp by using a high bending energy penalty to make large displacements of the forecast less likely to occur. In real situations, this works very well. However, as will be seen in section 4b, it is still possible for the image warp to allow large displacements if the displaced spatial structure in the forecast closely matches a (spatially distant) structure in the verification field.

The values of the average spatial displacement and reduction in error components of the IWS together determine how the weights \(c_{ij}\) should be chosen. When the average displacement is low, accompanied with a large reduction in error, then the combined contribution of \(D_j\) and \(1 - \eta_j\) should be larger than the contribution of \(\text{AMP}_j\). That is, \(c_{1j} + c_{2j}\) should be larger than \(c_{3j}\). If \(D_j\) and \(1 - \eta_j\) are low or medium in value, then their combined contribution should be about the same as the contribution of \(\text{AMP}_j\); that is, it is desired to have \(c_{1j} + c_{2j} \approx c_{3j}\). Finally, if the average displacement is high, then regardless of the amount of reduction in error, most users would want to ignore the spatial error information—relying solely on the original intensity error, \(\text{AMP}_j\), for information on forecast performance. In this last situation, it is desired to have \(c_{1j}\) and \(c_{2j}\) near zero and \(c_{3j}\) near one. It should be noted that in this situation, the image warp may attempt to squeeze together the false alarm areas to reduce their spatial extent and minimize their impact on the intensity error for the deformed forecast. Similarly, the image warp can *stretch* a small area of values in the forecast that occur near a larger area of values in the verification field (i.e., a missed area). Therefore, it may not be important to zero out the contribution of \(c_{1j}\) and \(c_{2j}\) for “real” cases, but for the perturbed cases of the ICP, it is more illustrative to set them to zero. Ultimately, as long as these general guidelines are followed, the ranking results are fairly insensitive to the exact values of these coefficients.

Lower values of IWS are better, and in its present form it is only meaningful for ranking multiple forecasts based on user-specific criteria. The issue is that the three components of the score have wildly different scales. One might come up with a reasonable normalization that
puts each term on equal footing, barring reweighting by the $c_p$ values, but this is beyond the scope of this paper. In this formulation, the displacement and amplitude components are standardized by subtracting their mean values and dividing by their standard deviations over the forecasts being compared.

3. Test cases

For the present study, we focus on QPFs, making use of the test cases from the ICP. The initial test cases comprise nine valid times from the National Oceanic and Atmospheric Administration/National Severe Storms Laboratory/Storm Prediction Center (NOAA/NSSL/SPC; information online at http://www.spc.noaa.gov/) 2005 Spring Program found to have interesting dynamics. Three configurations of the Weather Research and Forecasting (WRF; information online at http://www.wrf-model.org) model are compared with corresponding stage II analysis fields (i.e., used here as observations), all of which consist of 24-h accumulated precipitation interpolated to the same 601 $\times$ 501 grid with about 4-km resolution. Specifically, the models compared are the Nonhydrostatic Mesoscale Model (NMM), the Advanced Research WRF (ARW), and a model from the University of Oklahoma’s Center for Analysis and Prediction of Storms. [See Baldwin and Elmore (2005), Kain et al. (2008), Skamarock et al. (2005), and Davis et al. (2009) for more details on all of the ICP test cases with ease. Displacement errors are found within 12 grid points applied from points where precipitation occurs. For this reason, in subsequent analyses here, only vectors applied from points where precipitation occurs are used in calculating the average displacements.

We also demonstrate the technique on the perturbed WRF and contrived geometric cases from the ICP [see Ahijevych et al. (2009) for more details on all of the ICP test cases]. In addition to the ICP test cases, we also explore aggregating results across numerous cases using the NMM and ARW models for the entire 32-day period from the same 2005 Spring Program output; the same 32 cases were analyzed using the method for object-based diagnostic evaluation (MODE) in Davis et al. (2009).

4. Results and interpretations

We begin in section 4a with an analysis of the simple geometric cases, and then proceed in section 4b to scrutinize the method on the perturbed QPF examples; in particular, the IWS score for ranking multiple forecasts is put to the test. Section 4c applies the image warp to the real ICP test cases. Finally, in section 4d, we explore ways of aggregating information over multiple days using results from the image-warping scheme.

a. Analysis of geometric examples

The image-warping technique handles the geometric cases with ease. Displacement errors are found within a negligible amount of error when no penalty is applied to such translations. For all geometric cases, including the more complicated ones (e.g., overforecasting spatial extent and spatial displacement), the reduction in RMSE is nearly 100%. The only issue worth mentioning for these cases pertains to the fourth ICP geometric case (cf. Ahijevych et al. 2009) displayed here in the first two panels of Fig. 1.

In Fig. 1, the warp applies a rescaling to squeeze and stretch the forecast image instead of rotating it, as a human observer might be wont to doing. The main reason for this action by the image warp is that the area of higher intensity is asymmetrically placed to the right in the forecast. If the forecast had been a true rotation, the area of higher values would be placed in the center along the abscissa of the area with lower precipitation, and slightly up or down from the center along the ordinate axis (cf. with the observation). We applied the image-warping technique to a truly rotated forecast (not shown), and for this case, the resulting deformations represented a rotation. One might find this sensitivity alarming, but carefully chosen penalties in the prior (4) easily mediate this issue. Furthermore, regardless of the penalties chosen, we have found that this is not an issue with any of the real QPF cases tested.

Finally, the warp vector field shown in Fig. 1 (middle panel of bottom row) highlights the fact that the warp function is applied globally. It is based on the subset of control points used to determine the deformations. Therefore, there are vectors even where no precipitation values occur. For this reason, in subsequent analyses here, only vectors applied from points where precipitation occurs are used in calculating the average displacements.

b. Analysis of perturbed QPF examples

To obtain a better idea of how the various techniques work, the ICP has also provided cases that are simple perturbations of the model output so that the forecast in each case involves simple, known, errors such as spatial displacement. The image-warping technique handles these cases as one would expect; because the two images are very similar apart from deformations handled by the image-warp function. For example, the seventh of these ICP cases (pert007; see Table 1, Fig. 2) gives an optimal warp, after thresholding out the lowest 75% of values to remove bias and scatter, 1 that moves the forecast area on average to the west by just under 10 grid points (compared to the known displacement error of 12 grid

---

1 Section 4c explores the sensitivity of results to thresholding in terms of the amount of thresholding, and it is found that the results are fairly insensitive. Applying an absolute versus percentile threshold is more difficult to compare, but the choice also appears to be insensitive for these cases.
points too far east) and 24.28 grid points to the north (compared to the true displacement error of 20 grid points too far south). Finally, the reduction in error is about 60.9%, which is not so near to 100% as for most of the other perturbed cases because the intensities have also been reduced in the forecast. For this example, a high bending energy penalty is used, but results are similar for much lower penalties.

In fact, the sensitivity of the image warp was tested for each of the perturbed cases against a range of thresholds and bending energy penalties (not shown), and the results are substantially insensitive to these choices. This is perhaps a caveat of comparing against essentially the identical image with only very simple affine transformations. Therefore, if a forecast is very similar in spatial extent and shape, but displaced, then regardless of the bending energy penalty chosen, the image warp will pick up on the spatial displacement.

Table 1 shows the results from applying the image warp to the seven perturbed test cases. A useful exercise is to compare the performance of the seven cases. To do this, we employ the ranking algorithm described in section 2d. We choose our coefficients $c_{ij}$, $i = 1, 2, 3$ and $j = \text{pert001}, \ldots, \text{pert007}$, carefully to ensure that appropriate weight is given to each element of the score. For instance, the first case is a small perturbation (only an

![Fig. 1. Geometric case with the “forecast” spatially displaced 125 grid points to the right (east) and aspect ratio error.](image-url)
average displacement of $\sqrt{3^2 + 5^2} = 5.83$ grid points) with no amplitude discrepancies. For this case, the reduction in RMSE is relatively moderate compared to the other cases, but its original RMSE is also smaller. Therefore, per the general guidelines discussed in section 2d, we choose $c_{1,pert001} = c_{2,pert001} = c_{3,pert001} = \frac{1}{2}$.

Table 2 gives the choices of coefficients used for all analyses here for each combination of categories. The choice made is user dependent and is determined a priori. Following the guidelines described in section 2d, the weights $c_{ij}$ in the IWS vary depending on the amount of average displacement invoked by the image warp together with the amount of error reduction. Specific values of error reduction considered to be high or low are highly relative, and are best determined from the values obtained across the cases being compared. In this case, we use the upper and lower quartiles of the values for the cases considered. That is, a reduction in error is considered low if it is at or below the lower quartile (in this case 76.40%), high if it is above the upper quartile (here, 93.15%), and medium if it falls between these percentages. The amount of average displacement determining low, medium, and high amounts is more easily decided upon in advance. For example, one might consider that anything below 40 km$^2$ of average displacement is low, and anything above 400 km$^2$ is high. For illustrative purposes here, we again use the quartiles so that any average spatial displacement below 18 pixels (about 72 km) is considered low, and anything above 36 pixels (about 144 km) is considered high.

Subsequently, for the rankings derived in Table 1, pert001, pert006, and pert007 use $c_{1j} = c_{2j} = \frac{1}{4}$ and $c_{3j} = \frac{1}{2}$, as each has low or medium values for both of the first two components of IWS. The second case, pert002, uses

---

**Table 1.** Summary of seven ICP perturbed cases and image-warping results using a high bending energy penalty ($10^{4+1/3}$) and the 75th percentile threshold on the raw fields. Rank is determined here by the algorithm described in section 2d, and the actual (relative) IWS score is given in parentheses.

<table>
<thead>
<tr>
<th>Case</th>
<th>Description</th>
<th>Avg pixel movement</th>
<th>% reduction in RMSE ($\eta$)</th>
<th>RMSE (original)</th>
<th>Rank (IWS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>pert001</td>
<td>3 points right, 5 points down</td>
<td>7.56</td>
<td>82.9</td>
<td>1.71</td>
<td>1 ($-0.93$)</td>
</tr>
<tr>
<td>pert002</td>
<td>6 points right, 10 points down</td>
<td>13.10</td>
<td>96.9</td>
<td>2.03</td>
<td>2 ($-0.33$)</td>
</tr>
<tr>
<td>pert003</td>
<td>12 points right, 20 points down</td>
<td>24.59</td>
<td>95.0</td>
<td>2.27</td>
<td>3 ($-0.08$)</td>
</tr>
<tr>
<td>pert004</td>
<td>24 points right, 40 points down</td>
<td>48.02</td>
<td>91.3</td>
<td>2.32</td>
<td>5 (0.18)</td>
</tr>
<tr>
<td>pert005</td>
<td>48 points right, 80 points down</td>
<td>94.83</td>
<td>82.0</td>
<td>2.36</td>
<td>6 (0.29)</td>
</tr>
<tr>
<td>pert006</td>
<td>12 points right, 20 points down, $\times 1.5$</td>
<td>24.58</td>
<td>70.8</td>
<td>2.90</td>
<td>7 (0.89)</td>
</tr>
<tr>
<td>pert007</td>
<td>12 points right, 20 points down, $-1.27$ mm</td>
<td>25.62</td>
<td>60.9</td>
<td>2.19</td>
<td>4 ($-0.06$)</td>
</tr>
</tbody>
</table>

---

**Fig. 2.** Image-warp results for ICP case pert007 (see Table 1). The “forecast” is spatially displaced 12 grid points (about 48 km) too far to the east and 20 grid points (about 80 km) too far south, while the intensities have been reduced by 1.27 mm (0.05 in) from the truth field. Here, a threshold has been applied to both fields to leave out the lowest 75% of the values from each field (i.e., removing frequency bias and scatter). A bending energy penalty of $10^{4+1/3}$ was used to obtain this warp. (a) The original forecast field with vectors indicating the warp movements, (b) the verification field, and (c) the deformed forecast field.
To check the sensitivity of the image warp to choices of the bending energy penalty and threshold, the procedure is applied to various values of each. In general, the resulting values of average spatial displacement and reduction in error are relatively insensitive to these choices. For example, for the 4 June ICP test case, the average movement varies from about 34 to 38 grid squares, a discrepancy of about 16 km² in total displacement over the field. Further, the reduction in error only varies by about 2%. Inspection of the actual warps, however, reveals that a higher value for the bending energy penalty results in more easily interpretable displacements. Lower penalties allow for more complicating deformations, which may be of interest for some users, but too low of a penalty will often result in warps that are not physically meaningful (e.g., folding). Therefore, we select a relatively high bending energy penalty of about $2.15 \times 10^4$, and for the present study, we use the upper quartile for the threshold in order to remove excessive scatter, and maintain comparable results when comparing different models.

The warped forecast field in the right panel of Fig. 3 clearly matches the verification field better than the original field (lower-left panel), but unlike with the geometric or perturbed cases, the optimization routine stops before fitting very precisely so that the reduction in error (about 29%) is substantially less than for the those cases. An absolute threshold of 1.5 mm is also applied for the 4 June case. The reduction in error is less (about 26%), and the associated displacement is slightly greater (about four grid squares). It is difficult to make any meaningful comparisons between using an absolute versus relative threshold because it is not clear which absolute threshold best compares with which relative threshold. Therefore, we focus solely on using relative thresholds for the remainder of this study, which removes any spatial extent biases. Looking at results for all of the cases, however, this does not have an impact on the results.

Figure 4 compares the IWS ranks against the subjective score ranks. Perfect agreement would produce a plot with a perfectly straight diagonal line from the lower-left corner of the plot to the upper-right corner. Clearly, this is not the case for either model. Of course, the subjective scores (cf. Ahijevych et al. 2009) are averaged over fewer than 30 individuals, with much

---

**Table 2.** Choices for coefficients used to create rankings of the seven perturbed ICP cases using the IWS (6) in Table 1. Category combinations not shown have equally weighted coefficients (i.e., $c_{1j} = c_{2j} = c_{3j} = 1/3$).

<table>
<thead>
<tr>
<th>Category</th>
<th>Avg displacement ($D_j$)</th>
<th>Reduction in error ($\eta_j$)</th>
<th>Amplitude (AMP)</th>
<th>Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>Low</td>
<td>Any</td>
<td>3/8, 3/8, 1/4</td>
<td></td>
</tr>
<tr>
<td>Low or medium</td>
<td>Low or medium</td>
<td>Any</td>
<td>1/4, 1/4, 1/2</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>Any</td>
<td>0, 0, 1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$c_{1\text{pert002}} = c_{2\text{pert002}} = 3/8$, and $c_{3\text{pert002}} = 1/4$ because it has low displacement and high reduction in error relative to the other choices. Pert003 uses equal weights (i.e., $c_{1\text{pert003}} = c_{2\text{pert003}} = c_{3\text{pert003}} = 1/3$), having medium displacement with high reduction in error. Finally, pert004 and pert005 both use $c_{1j} = c_{2j} = 0$ and $c_{3j} = 1$ because of their relatively high displacement errors.

The weights make sense in terms of physical meaning. For instance, because an extremely high displacement is needed to match the forecast fields for pert004 and pert005 to the observed field, the errors should arguably be counted as false alarms and misses rather than displacement errors so that the original RMSE value is correctly given all of the weight. Generally, the rankings are found to be fairly insensitive to the choices of weights as long as the relative weighting is consistent. Despite the lower RMSE of pert007 compared with pert003, the two are displaced by the same amount. When this displacement is corrected, of course, pert003 obtains significant improvement in RMSE (95.0% compared with only 60.9%). Therefore, it can be argued that the rankings correctly place pert003 ahead of pert004.

Of course, from inspection of the IWS components in Table 1, it is clear that values are fairly close across cases. Therefore, one might question whether the differences in IWS are substantial enough or not to inspire confidence in the rankings. Without confidence intervals (or a hypothesis test) for the differences in IWS between the cases, it is not possible to make claims on the statistical significance of the difference. Consistent rankings (e.g., forecast A is always ranked ahead of forecast B) might indicate practical significance. However, assuming normality for the differences in IWS scores creates confidence intervals that are perhaps too conservative; only very large spatial errors are found to be significant, such as between pert001 and pert005 at the 5% level. Therefore, we do not attempt to discuss confidence intervals for these ranks further, and merely utilize the IWS to assist in summarizing the image-warping results.

**c. Nine real ICP cases**

To check the sensitivity of the image warp to choices of the bending energy penalty and threshold, the procedure is applied to various values of each. In general, the resulting values of average spatial displacement and reduction in error are relatively insensitive to these choices. For example, for the 4 June ICP test case, the average movement varies from about 34 to 38 grid squares, a discrepancy of about 16 km² in total displacement over the field. Further, the reduction in error only varies by about 2%. Inspection of the actual warps, however, reveals that a higher value for the bending energy penalty results in more easily interpretable displacements. Lower penalties allow for more complicating deformations, which may be of interest for some users, but too low of a penalty will often result in warps that are not physically meaningful (e.g., folding). Therefore, we select a relatively high bending energy penalty of about $2.15 \times 10^4$, and for the present study, we use the upper quartile for the threshold in order to remove excessive scatter, and maintain comparable results when comparing different models.

The warped forecast field in the right panel of Fig. 3 clearly matches the verification field better than the original field (lower-left panel), but unlike with the geometric or perturbed cases, the optimization routine stops before fitting very precisely so that the reduction in error (about 29%) is substantially less than for the those cases. An absolute threshold of 1.5 mm is also applied for the 4 June case. The reduction in error is less (about 26%), and the associated displacement is slightly greater (about four grid squares). It is difficult to make any meaningful comparisons between using an absolute versus relative threshold because it is not clear which absolute threshold best compares with which relative threshold. Therefore, we focus solely on using relative thresholds for the remainder of this study, which removes any spatial extent biases. Looking at results for all of the cases, however, this does not have an impact on the results.

Figure 4 compares the IWS ranks against the subjective score ranks. Perfect agreement would produce a plot with a perfectly straight diagonal line from the lower-left corner of the plot to the upper-right corner. Clearly, this is not the case for either model. Of course, the subjective scores (cf. Ahijevych et al. 2009) are averaged over fewer than 30 individuals, with much
variability (e.g., different individuals may score a forecast based on different criteria). Further, some of the scores are very close together so that differences in ranking multiple close scores should not be considered substantial. Therefore, one should not look at the subjective scores as necessarily being the truth, or their rankings as a definitive ordering. Indeed, a more carefully planned subjective evaluation would be necessary to obtain more reliable information about how well the image warp matches with human observation. However, this is beyond the scope of the present paper. Nevertheless, the correspondences between the two rankings can give some indication of how the image-warp method performs; particularly in terms of major differences in the ranks.

It would seem that the subjective ranks agree better with the IWS ranks for the NMM model, where the IWS and subjective judgment both roughly agree on which cases are the best and worst. The only major discrepancies being that the fourth best case for NMM corresponds with the penultimate case according to the IWS rankings, and the best NMM for the IWS ranked fifth subjectively. For the ARW model, the two ranking systems disagree about the best and worst cases, where the top IWS case ranked last subjectively, and the two best ARW cases according to subjective opinion ranked last according to the IWS. Apart from these cases, however, the rest of the ARW ranks agree very well.

The largest disagreements between the two ranking systems are summarized in Table 3. When ranking the nine cases based on RMSE (not shown), the IWS rankings generally do not agree well at all for the NMM model. However, the two poorest matches with subjective evaluation rankings (19 May and 1 June) match almost perfectly with those for the RMSE ranks. Inspection of the 19 May case (see Ahijevych et al. 2009) shows that there is relatively little precipitation observed, and it is possible that the subjective evaluators may have put a lot of emphasis on low-intensity scatter that was removed via thresholding in implementing the image-warp method here. For the 1 June case, it appears that the overall structure of the storm system was reasonably

![Image](image-url)
Table 3. Worst-rank matchings between the IWS ranking method and the subjective evaluation rankings from having ranked the nine real ICP test cases for each of the ARW and NMM models. Ranks based on RMSE are shown for comparison.

<table>
<thead>
<tr>
<th>Case</th>
<th>Model(s)</th>
<th>IWS rank</th>
<th>Subjective rank</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>13 May</td>
<td>ARW</td>
<td>8</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>19 May</td>
<td>ARW/NMM</td>
<td>1/1</td>
<td>9/5</td>
<td>3/1</td>
</tr>
<tr>
<td>1 June</td>
<td>ARW/NMM</td>
<td>9/8</td>
<td>2/4</td>
<td>9/9</td>
</tr>
</tbody>
</table>

well matched by the NMM, but that it is generally displaced a bit too far to the west, with much greater intensities (hence the poor RMSE ranking).

Comparison between the ranks and IWS ranks for the ARW model suggest that there is better agreement. For the worst matches with subjective evaluation (13 May, 19 May, and 1 June), the IWS rank again agrees very well with the RMSE rankings. For the 13 May case, it appears that the ARW model has about the correct structure and relatively good placement, but perhaps overestimates the intensity over a wide area to account for the low relative rank for both the IWS and RMSE. On 19 May, it appears that the ARW model again has too much intensity, but the image warp did not require much energy to morph the field to better match the observed field. In this case, the lack of precipitation is perhaps responsible for the higher ranking and suggests that one might want to consider the total area of precipitation observed when deciding on weights for the IWS. However, because this pattern of behavior depends on the observed field, it is not possible to simply hedge the forecast to improve the ranking. Similar to the NMM model, for the 1 June case, it appears that the ARW model had roughly the right spatial extent and structure of the storm system, but is displaced some, and overpredicts the intensities. This may account for the discrepancy in ranking against subjective evaluation, as well as the agreement with RMSE ranking for this case.

d. Aggregating over multiple days

We have demonstrated that the image warp provides intuitive, sensible information when applied to cases with known errors, as well as for real cases where some notion of expert analysis provides guidance for testing the method. Indeed, the image-warp procedure provides detailed diagnostic information when forecast performance can be conducted case by case, and studied in detail. Often, however, information over multiple fields is needed. In such a case, it may not be possible to scrutinize each situation for specific forecast errors. Therefore, summary information can be useful, and the image warp can be used in this way while being spatially informed. Here, results are aggregated over the 32 Spring Program cases, and compared with previous analysis.

A bending energy penalty based on quilt plots for the case results in section 4c of about $2.15 \times 10^4$ is used. Additionally, the upper quartile for the relative threshold is imposed.

Results indicate that the ARW model is generally slightly better than the NMM model, but this is largely a result of a few days. The day-to-day performance shows much variability, and most days are indistinguishable in terms of performance. For example, both models overpredicted the spatial extent of the precipitation events for 10 and 11 May, but the spatial extent bias for NMM was greater in both cases, and the reduction in RMSE for either is relatively low. For most days of the period, the ARW model scored better (i.e., lower) in terms of the IWS than NMM (Fig. 5, top). Figure 5 (second row) shows the amount of spatial displacement invoked by the image warp along with the associated reduction in error (third row). Both forecasts are displaced substantially from about 27 April until about 5 May, and the resulting reduction in RMSE is also substantially greater during this time period. However, the original RMSE (Fig. 5, bottom) was higher for several dates later in the spring. Without inspecting the individual cases, it is impossible to decipher this behavior. Inspection of each case, however, reveals that there are larger areas of observed precipitation in the early part of the spring, whereas the later parts are marked by very sparse observed precipitation areas (not shown). These results are in general agreement with those found by Davis et al. (2009), who used aggregated statistics from MODE to analyze these same 32 cases.

5. Summary, conclusions, and discussion

The image warp is applied and tested on the ICP test cases. The technique performs extremely well for the simple, but known, errors of the geometric and perturbed cases; giving nearly exact information about the forecast’s spatial errors. When applied to realistic cases, imposed constraints (e.g., limited numbers of control points and high bending energy penalties) result in warped fields that do not match the observed fields as precisely, but nevertheless provide sensible and useful information about forecast performance.

A statistic is proposed for ranking multiple forecasts based on their resulting image warps. The technique requires user-specified parameters that can affect the outcome of the rankings, but results are reasonably insensitive to these choices. For example, large average displacements can be up- or down-weighted depending on specific user needs in interpreting such errors. In the case of the IWS, the ARW model had roughly the right spatial extent and structure and relatively good placement, but perhaps overestimates the intensity over a wide area to account for the low relative rank for both the IWS and RMSE.
applied to the perturbed ICP cases revealed that the IWS ranking algorithm ranks the forecasts sensibly. The exercise draws attention to the user-specific question as to whether pert007 (cf. Ahijevych et al. 2009) is better or worse than pert003. The two have identical displacement errors, but the former also has less intensity. Therefore, if seen as a pure displacement, pert003 is clearly better, but if seen as a true false alarm (and miss), then pert007 is better as it has a less severe false alarm. Because the displacement error for these cases is relatively small, the IWS ranking correctly slots pert003 ahead of pert007. Nevertheless, choices in parameters for this statistic can result in ranking pert007 higher, but such choices are physically meaningful and are otherwise insensitive to small changes in their values.

Applying the IWS ranking procedure to the nine real ICP cases and comparing them with subjective ranks reveals that the method generally agrees well with subjective assessment. However, some stark differences occur. In these cases, it may be argued that the subjective ranks may have been too heavily influenced by one type of positive (or negative) feature. For example, the 1 June case ranked highly among subjective evaluators probably because the models captured the general storm structure fairly well. However, in addition to a small spatial displacement, the models greatly overpredicted the intensities.

When applied to the extended set of 2005 Spring Program cases, the results indicate that the ARW model generally faired better than the NMM model, but that this overall finding is the result of relatively few cases (there is much day-to-day variability in their performances). These results are in accordance with Davis et al. (2009), who applied MODE, another spatial forecast verification method, to these test cases.

For any given type of field, some care is required in setting up and testing the image-warp procedure. Once an appropriate likelihood and prior distribution are determined for a particular type of field, however, the method works very well in terms of diagnosing specific forecast errors. The method is also naturally formulated through a solid statistical framework (cf. Murphy and Winkler 1987), as is evident by model (1). Through this model, it is possible to extend the principle to incorporate more advanced information (e.g., more dimensions such as temporal and/or vertical components,
multiple fields, etc.). Such additions may increase the computational effort, and careful thought in formulating the likelihood/prior is necessary. The current implementation of the image-warping method can easily be used in an operational setting.

We have shown here how the image-warping method can be used for precipitation fields, but initial results for other fields suggests that it should be useful for them as well. In particular, the success with handling the geometric cases suggests that it might work well in providing additional attributes for the MODE tool (Davis et al. 2009).

All of the image warps in this study used a preset, sparse, regular grid for the control points. Given that the resulting warps for all of the cases studied provided sound information about forecast error, this study demonstrates an important advancement in the use of image warps for forecast verification because past studies were heavily concerned about identifying features. It is found here that this is not necessary in order to glean the necessary information about forecast performance.

Acknowledgments. The authors would like to acknowledge Swedish Foundation for International Cooperation in Research and Higher Education (STINT) Grant IG2005-2047, which made this collaboration possible. This work was sponsored in part by the National Center for Atmospheric Research (NCAR).

REFERENCES


