NCAR/TN-482+STR
NCAR TECHNICAL NOTE

2010-10-26

Spatial Forecast Verification: Image Warping

Eric Gilleland, Linchao Chen, Michael DePersio, Giang Do, Kirsten Eilertson, Yin Jin, Emily L. Kang, F. Lindgren, J. Lindström, Richard L. Smith, and Changming Xia

Institute for Mathematics Applied to Geosciences (IMAGe) and
Research Applications Laboratory

NATIONAL CENTER FOR ATMOSPHERIC RESEARCH
P. O. Box 3000
BOULDER, COLORADO 80307-3000
ISSN Print Edition 2153-2397
ISSN Electronic Edition 2153-2400

NCAR TECHNICAL NOTES
http://www.ucar.edu/library/collections/technotes/technotes.jsp

The Technical Notes series provides an outlet for a variety of NCAR Manuscripts that contribute in specialized ways to the body of scientific knowledge but that are not suitable for journal, monograph, or book publication. Reports in this series are issued by the NCAR scientific divisions. Designation symbols for the series include:

EDD – Engineering, Design, or Development Reports
Equipment descriptions, test results, instrumentation, and operating and maintenance manuals.

IA – Instructional Aids
Instruction manuals, bibliographies, film supplements, and other research or instructional aids.

PPR – Program Progress Reports
Field program reports, interim and working reports, survey reports, and plans for experiments.

PROC – Proceedings
Documentation or symposia, colloquia, conferences, workshops, and lectures. (Distribution may be limited to attendees).

STR – Scientific and Technical Reports
Data compilations, theoretical and numerical investigations, and experimental results.

The National Center for Atmospheric Research (NCAR) is operated by the nonprofit University Corporation for Atmospheric Research (UCAR) under the sponsorship of the National Science Foundation. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

NCAR/TN-482+STR
NCAR TECHNICAL NOTE

Spatial Forecast Verification: Image Warping

Eric Gilleland
Joint Numerical Testbed, Research Applications Laboratory, National Center for Atmospheric Research

Linchao Chen
Ohio State University

Michael DePersio
Georgetown University

Giang Do
Georgia Institute of Technology

Kirsten Eilertson
Cornell University

Yin Jin
University of California, Davis

Emily L. Kang
Statistical and Applied Mathematical Sciences Institute and North Carolina State University, Dept. of Statistics

Finn Lindgren
Norwegian University of Science and Technology, Trondheim, Norway

Johan Lindström
Mathematical Statistics, Centre for Mathematical Sciences, Lund University, Lund, Sweden

Richard L. Smith
Statistical and Applied Mathematical Sciences Institute and University of North Carolina, Chapel Hill

Changming Xia
University of Rochester Medical Center

NATIONAL CENTER FOR ATMOSPHERIC RESEARCH
P. O. Box 3000
BOULDER, COLORADO 80307-3000

Contents

List of Figures

List of Tables

PREFACE

In response to a growing need for more informative forecast verification in the face of gridded verification sets, many new methods have been proposed. While widely varying in their approaches, the new methods generally fall into two major categories of filter and displacement, each of which can be further subdivided. One of the displacement approaches, a field deformation approach known as image warping, will be demonstrated here. Results for spatial verification of the spatial forecast verification Inter-Comparison Project test cases are shown. An initial look at space-time verification using the image warp is also discussed, with an application to NCAR and NCEP 4-km WRF models cases from the 2005 NSSL/SPC Spring Program. The approach is found to be very useful for obtaining guidance about forecast performance. Both diagnostic and summary score information can be gleaned. Initial findings for the space-time approach show that while the NCEP model has better initial scores, the NCAR models require drastically less deformation to achieve a much higher reduction in error. This is most likely a result of the NCEP model's highly over forecasting low-intensity precipitation spatially.

1 Introduction

Spatial forecast verification has received a lot of attention in the last decade largely because of the advent of higher resolution models, which present difficulties for traditional grid-point to grid-point verification (see e.g., Mass et al., 2002). While high-resolution forecast models may provide forecasts for a particular variable, it may be difficult to find good observational data with which to make comparisons that are not generated by the same model (e.g., at shorter lead times). Data assimilation is one way to avoid both problems, but it is often of interest to make comparisons with gridded observations. Here, it is assumed that the verification set, which contains two fields: an observed field and a forecast field, is gridded with both fields on the same grid.

Gilleland et al. (2009) categorized the various new methods into one of four categories, whereby some methods fit only loosely: (i) features-based, (ii) field deformation, (iii) neighborhood, and (iv) scale separation. The first two can be considered as *displacement* methods, and the latter two as *filter* methods. Examples of features-based methods include Ebert and McBride (2000); Nachamkin (2004, 2009); Davis et al. (2006a,b, 2009); Grams et al. (2006); Marzban and Sandgathe (2006a,b); Wernli et al. (2008, 2009); Ebert and Gallus (2009), where the methods of Nachamkin (2004, 2009) and Marzban and Sandgathe (2006a,b) fit less well into the category. Examples of neighborhood approaches are numerous, but Ebert (2008) provides a thorough review. Fourier decomposition is a scale separation method that has long been applied, but other examples include wavelet decomposition (e.g., Briggs and Levine, 1997; Casati et al., 2004; Casati, 2009), and structure function type approaches, which fit less well (e.g., Harris et al., 2001; Marzban and Sandgathe, 2009). Field deformation approaches are among the earliest methods to be proposed (e.g., Hoffman et al., 1995; Alexander et al., 1999; Nehrkorn et al., 2003), but new attention has been given to them recently (e.g., Keil and Craig, 2007, 2009; Gilleland et al., 2010b). Most of these methods have been analyzed for quantitative precipitation fields (QPFs) as part of the spatial forecast verification inter-comparison project (ICP, http://www.ral.ucar.edu/projects/icp).

Field deformation methods are techniques applied to an entire field that attempt to quantify location errors including spatial pattern errors. The simplest techniques are mathematically defined image metrics (e.g., Venugopal et al., 2005; Gilleland et al., 2008), which simply give a summary statistic describing the closeness of the spatial patterns. More ambitious field deformation approaches attempt to morph the forecast field (and/or the observed field) to better match its counterpart. There are vastly many ways to deform a field. Image warping is of interest here because of its statistical elegance relative to other such methods.

There are many references pertinent to image warping and spatial statistics, but the reader is directed to Åberg et al. (2005) and Gilleland et al. (2010b) as

they correspond most closely with the techniques described herein, and further references can be found in these articles.

There are numerous possible extensions to the image warping described in this document that could prove very useful to forecast verification. For example, inclusion of covariate information might yield more meaningful results (e.g., if topography could be included to allow for more physically meaningful warps). Adding a vertical dimension could also be useful if information for different vertical layers are available. This could, for example, answer the question of how well a forecast is predicting activity at the correct pressure levels.

2 The Image Warp

The basic idea of using image warping for forecast verification is simple. Deform the forecast field spatially so that the intensities align better with the observed intensities. Information about how much deformation is needed, and how much reduction in error as determined by traditional grid-point to grid-point verification (e.g., rmse) can be used to analyze forecast performance.

2.1 The Model

When using the image warp for forecast verification purposes, the model is

$$O(x, y) = F(x', y') + \varepsilon(x, y), \tag{1}$$

where O is the observed field, F the forecast field, ε are intensity errors, and x and y are coordinates. Note that the model is assuming that the observed field at any given coordinate (x, y) is approximately equal to the forecast field at possibly different coordinates (x', y'). Image warping is essentially a smooth mapping between two image planes, in this case an observed and forecast field, that maps every point in one field to a point in the other. To deform an image, every point (x', y') in the deformed image is assigned the intensity of the undeformed image in the point with coordinates given by $(W_x(x, y), W_y(x, y))$, where W is a mapping of points (x, y) from O to F. It is useful to write the image warping function, W, as a vector, $\boldsymbol{W}(x, y) = (W_x(x, y), W_y(x, y))^T$.

Subsequently, the model (1) can be formulated as

$$\begin{aligned} \widehat{O}(x, y) = \widetilde{F}(x, y) &= F(W_x(x, y), W_y(x, y)) \\ &= F(\boldsymbol{W}(x, y)), \text{ for all } (x, y) \in \mathcal{D}, \end{aligned} \tag{2}$$

where $\mathcal{D}$ is the support of the field (i.e., the grid domain). Note that the mapping, $\boldsymbol{W}$, is from coordinates in the estimate of the observed field (i.e., the deformed

forecast field, $\widetilde{F}$) to coordinates in the forecast field. In order to obtain the deformed forecast field each grid point (x, y) in $\widetilde{F}$ is found by identifying the relevant grid point value $\boldsymbol{W}(x, y)$ in F. A couple of issues arise because some grid points might come from outside the domain of F, or they might not coincide with the middle of the grid point exactly. Section 3 discusses possible ways to handle these issues.

2.2 The Likelihood

The deformation induced through image warping is entirely determined by the grid point mapping $\boldsymbol{W}$. For the thin-plate spline warp (see Section 2.3) used here, the function $\boldsymbol{W}$ depends only on how a fixed set of points are mapped. Therefore, we fix m points $\boldsymbol{p}^O = \{\boldsymbol{p}_1^O, \ldots, \boldsymbol{p}_m^O\}$ in the observed field. Here, we will refer to these points as *control* points; other names have been used such as tie points, landmarks, etc. In early papers applying image warping to forecast verification, attempts were made to identify features within the fields in order to determine where these points should be located. However, this is not necessary to obtain meaningful deformations. For example, Gilleland et al. (2010b) show that informative deformations are obtained using a regular grid of control points, which is the approach used here.

Next, define $\boldsymbol{p}^F = \{(\boldsymbol{W}(\boldsymbol{p}_1^O), \ldots, (\boldsymbol{W}(\boldsymbol{p}_m^O))\} = \boldsymbol{W}(\boldsymbol{p}^O)$, as the maps of the observed control points in the forecast field, called the forecast control points. The deformed field, say $\widetilde{F}$, is given by $\widetilde{F}(\boldsymbol{s}) = F(\boldsymbol{W}(\boldsymbol{s}))$. To measure the dissimilarity between $\widetilde{F}$ and O, an error log-likelihood for $O - \widetilde{F}$ is introduced.

$$\log \mathrm{p}(O|F, \boldsymbol{p}^F, \boldsymbol{p}^O, \boldsymbol{\theta}) = f(O, \widetilde{F}; \boldsymbol{\theta}). \tag{3}$$

The image warp procedure can be thought of as taking the forecast grid as though it were a piece of paper, shift it about (e.g., translate it up or down, rotate it, crinkle it up, etc.). If left alone, optimization of the log-likelihood in (3) may well lead to non-physical warps (e.g., imagine folding the paper over on itself) that would be difficult to interpret for forecast verification purposes. Therefore, standard practice is to introduce a prior df in order to penalize such non-physical warps. Because the warp is entirely determined by the movements of the control points, this amounts to a prior on the control points, but because the observed control points are fixed *a priori*, this reduces further to $\mathrm{p}(\boldsymbol{p}^F|\boldsymbol{p}^O, \boldsymbol{\theta})$. Allowing for a prior df on the parameters $\boldsymbol{\theta}$ of the assumed df for the errors (i.e., $\mathrm{p}(\boldsymbol{\theta})$), the posterior log-likelihood becomes

$$\ell(\boldsymbol{p}^F, \boldsymbol{\theta}) = \log \mathrm{p}(O|F, \boldsymbol{p}^F, \boldsymbol{p}^O, \boldsymbol{\theta}) + \log \mathrm{p}(\boldsymbol{p}^F|\boldsymbol{p}^O, \boldsymbol{\theta}) + \log \mathrm{p}(\boldsymbol{\theta}) + \text{constant}, \tag{4}$$

The exact form of the log-likelihood in (4) is determined by the verification sets being applied. For the procedures described here, the penalties are chosen

and fixed so that the prior on the warp function reduces to $\log \mathrm{p}(\boldsymbol{p}^F|\boldsymbol{p}^O)$. A relatively simple, and generally valid, approach is described in section 2.4.

2.3 The Warp Function

There are numerous functions that can be used for image warping (e.g., polynomial, Procrustes, B-spline, thin-plate splines, etc.,). The thin-plate spline (TPS) is a good choice generally, and is described well in Åberg et al. (2005). Glasbey and Mardia (1998) give a thorough review of this and other image warping functions as well.

Note that the warp function chosen comes into play in determining the closeness of intensities between the observed field and the deformed forecast field through optimization of (4).

A TPS is a generalization of a cubic smoothing spline. It is a function, say f, that minimizes a two-dimensional Laplaceian penalty, or *bending energy*, given by

$$\mathcal{J}(f) = \int\int \left\{ \left(\frac{\partial^2 f}{\partial x^2}\right)^2 + 2\left(\frac{\partial^2 f}{\partial x \partial y}\right)^2 + \left(\frac{\partial^2 f}{\partial y^2}\right)^2 \right\} dxdy. \tag{5}$$

The solution to (5) has the form of a linear combination of radial basis functions. In the context of the warping function, this is applied to both the x- and y-coordinates. Specifically,

$$W_x(x,y) = a_0 + a_1 x + a_2 y + \sum_{i=1}^{m} d_{xi} U(\|\boldsymbol{p}_i^O - (x,y)\|),$$

where $U(r) = r^2 \log r$, $\sum_{i=1}^m d_{xi} = \sum_{i=1}^m d_{xi}x_i = \sum_{i=1}^m d_{xi}y_i = 0$. Here, the differences between (x,y) and the m control points are shown to emphasize that the TPS is fit only to the control points (but can be applied to every point in the domain to obtain $\widetilde{F}$). The function W consists of an affine part (rigid transformations) and a nonlinear part (the radial basis functions, U). It is possible to report these types of errors separately, which may be of interest as, for example, east-west displacements are summarized by a_x and north-south displacements by a_y, and combinations of these displacements can inform about scaling and rotation errors. While more difficult to interpret, the nonlinear part can give information about more intricate displacements. In this case, the amount of nonlinear displacement is a useful summary.

The above equation, with the various constraints can be more readily interpreted as a linear system of equations. Specifically,

$$\begin{bmatrix} \boldsymbol{K} & \boldsymbol{P} \\ \boldsymbol{P}^T & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{d}_x & \boldsymbol{d}_y \\ \boldsymbol{a}_x & \boldsymbol{a}_y \end{bmatrix} = \begin{bmatrix} \boldsymbol{p}_x^F & \boldsymbol{p}_y^F \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix},$$

where $[\boldsymbol{K}]_{ij} = U(\|\boldsymbol{p}_i^O - \boldsymbol{p}_j^O\|)$, the i-th row of $\boldsymbol{P}$ is $(1, x_i, y_i)$, $\boldsymbol{a}_x = (a_0, a_1, a_2)^T$ the affine coefficients for the x- coordinate (similar for $\boldsymbol{a}_y$), and $\boldsymbol{d}_x = (d_{x1}, \ldots, d_{xm})^T$ (similar for $\boldsymbol{d}_y$). Letting $\boldsymbol{L}$ denote the left most matrix on the left-hand side of the above equation, the minimum bending energy from Eq (5) is given by

$$\mathcal{J}(\boldsymbol{W}) = \mathcal{J}(W_x) + \mathcal{J}(W_y) = \boldsymbol{p}_x^T \boldsymbol{L}_m^{-1} \boldsymbol{p}_x + \boldsymbol{p}_y^T L_m^{-1} \boldsymbol{p}_y, \tag{6}$$

where L_m^{-1} is the upper left $m \times m$ subblock of $\boldsymbol{L}^{-1}$.

The extension of the TPS concept to three dimensions is known as a triharmonic spline, and the energy functional needing to be optimized is given by

$$\int\int\int \left\{ \left(\frac{\partial^2 f}{\partial x^2}\right)^2 + \left(\frac{\partial^2 f}{\partial y^2}\right)^2 + \left(\frac{\partial^2 f}{\partial t^2}\right)^2 + 2\left(\frac{\partial^2 f}{\partial x \partial y}\right)^2 + 2\left(\frac{\partial^2 f}{\partial x \partial t}\right)^2 + 2\left(\frac{\partial^2 f}{\partial y \partial t}\right)^2 \right\} dx dy dt.$$

Now, the warping function becomes

$$W(\boldsymbol{s}, t) = f(\boldsymbol{s}, t) + \sum_{i=1}^{n} d_i V(\|\boldsymbol{p}_i^O - (\boldsymbol{s}, t)\|),$$

where $V(r) = r^3$, $f(\boldsymbol{s}, t)$ is a polynomial of degree at most 2 (here we use degree one), and $\sum_i d_i = \sum_i d_i x_i = \sum_i d_i y_i = \sum_i d_i t_i = 0$. This leads to the linear system of equations

$$\begin{bmatrix} \boldsymbol{K} & \boldsymbol{P} \\ \boldsymbol{P}^T & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{d}_x & \boldsymbol{d}_y & \boldsymbol{d}_t \\ \boldsymbol{a}_x & \boldsymbol{a}_y & \boldsymbol{a}_t \end{bmatrix} = \begin{bmatrix} \boldsymbol{p}_x^F & \boldsymbol{p}_y^F & \boldsymbol{p}_t^F \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix},$$

where $[\boldsymbol{K}]_{ij} = V(\|\boldsymbol{p}_i^O - \boldsymbol{p}_j^O\|)$, and the rest as before except with the added time dimension.

2.4 Specific Models

As mentioned earlier, the normal assumption for the error likelihood (4) is generally reasonable. However, special considerations are required for precipitation fields. One way of how this can be handled is given at the end of the section.

For many fields, it is reasonable to assume that the errors are normally distributed. That is, $[O|F, \boldsymbol{p}^F, \boldsymbol{p}^O] \sim N(\widetilde{F}, \sigma_\varepsilon^2)$. For the prior df for the forecast control points given the observed ones, it is also common to assume a normal df. For example, following Åberg et al. (2005)

$$\begin{aligned} [\boldsymbol{p}_i^O - \boldsymbol{p}_i^F | \boldsymbol{p}_j^O - \boldsymbol{p}_j^F, j \neq i, F] &= [\boldsymbol{p}_i^O - \boldsymbol{p}_i^F | \boldsymbol{p}_j^O - \boldsymbol{p}_j^F, j \neq i] \\ &\sim N\left(\sum_{j \neq i} c_{ij} (\boldsymbol{p}^O - \boldsymbol{p}^F),\, \sigma_\Delta^2 \boldsymbol{I} \right) \end{aligned} \tag{7}$$

where $\boldsymbol{I}$ is a 2×2 identity matrix, indicating that the two coordinates are assumed to be independent, c_{ij} are scalars that specify the degree of dependence between control points i and j with $c_{ij} \neq 0$ if and only if grid points i and j are neighbors. Note that it is assumed here that the displacement of the control points are conditionally independent of the forecast field. If one has additional knowledge about how control points should move based on only one image (e.g., using topography as a covariate), then it is not necessary to assume such independence; but it is a reasonable, and simplifying, assumption here. Recalling that the warp is applied to both coordinates (using x and y subscripts on the control points to denote the corresponding displacement), df (7) leads to the joint df for the displacements given by

$$\begin{bmatrix} \boldsymbol{p}_x^O - \boldsymbol{p}_x^F \\ \boldsymbol{p}_y^O - \boldsymbol{p}_y^F \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_\Delta^2 (\boldsymbol{I} - \boldsymbol{C})^{-1} & 0 \\ 0 & \sigma_\Delta^2 (\boldsymbol{I} - \boldsymbol{C})^{-1} \end{bmatrix} \right) \tag{8}$$

where $[\boldsymbol{p}_x^O - \boldsymbol{p}_x^F, \boldsymbol{p}_y^O - \boldsymbol{p}_y^F]^T$ are the $x-$ and $y-$ coordinates stacked on each other in a column, and

$$\{\boldsymbol{C}\}_{ij} = \begin{cases} 0, & i = j \\ c_{ij}, & i \neq j \end{cases}$$

The interpretation of this prior model for the displacements is that the control points move in a similar way as their neighbors; thereby ensuring a smooth deformation, as well as penalizing for non-physical warps.

Under the assumptions above, and leaving out the constant parts that only depend on the fixed parameters σ_ε^2, σ_Δ^2 and $\boldsymbol{C}$, the following loss function that minimizes the error log-likelihood (4) becomes

$$\begin{aligned} &Q(\boldsymbol{p}^F) = \tfrac{1}{2\sigma_\varepsilon^2} \left(\sum_{i=1}^{n} (O(\boldsymbol{s}_i) - F(\boldsymbol{W}(\boldsymbol{s}_i)))^2 \right) + \\ &\tfrac{1}{2\sigma_\Delta^2} \left((\boldsymbol{p}_x^F - \boldsymbol{p}_x^O)^T (\boldsymbol{I} - \boldsymbol{C})(\boldsymbol{p}_x^F - \boldsymbol{p}_x^O) + (\boldsymbol{p}_y^F - \boldsymbol{p}_y^O)(\boldsymbol{I} - \boldsymbol{C})(\boldsymbol{p}_y^F - \boldsymbol{p}_y^O) \right). \end{aligned} \tag{9}$$

The first term penalizes differences between the observed and deformed forecast field intensities. Note that this term involves all points in the domain, but that the warp function itself is determined by only the subset of control points. The second term, which is applied only to the control points, penalizes for too large displacements and neighboring control points that move in different directions. The ratio $\sigma_\varepsilon^2/\sigma_\Delta^2$ specifies how much the prior will penalize certain mappings relative to the likelihood, and $\boldsymbol{C}$ defines the dependence between the control points.

The optimization of (9) must be performed numerically. For example, by using a robust Newton method. This, and other techniques, requires calculation of first derivatives. See Åberg et al. (2005) for specific details.

The above approach uses a Gaussian Markov random field approach for the prior on control point movements (i.e., $\boldsymbol{p}^O - \boldsymbol{p}^F$) because $\boldsymbol{C}$ is zero for points that are not *close* to each other. It is used here for the space-time extension because of its simplicity and it has the nice interpretation that the control points move in a similar way neighboring points. An alternative prior, used by Gilleland et al. (2010b), is the bending energy of the TPS. The bending energy is also a Gaussian prior, and amounts to replacing $\sigma_\Delta^2(\boldsymbol{I}-\boldsymbol{C})^{-1}$ in (8) with the bending energy matrix from the TPS (see Eq (6) above). Similarly for the inverse $(1/\sigma_\Delta^2(\boldsymbol{I}-\boldsymbol{C}))$ in Eq (9). The bending energy leads to an improper Gaussian prior since it does not penalize the affine part of the warp (translation, rotation, scaling and skew); only the nonlinear part is penalized.

For precipitation fields, the normal assumption for the errors $O-\widetilde{F}$ is generally not valid. The non-normality essentially consists of two parts: (i) dry weather at several locations leads to a zero-inflated error distribution (simultaneously no observed and no forecast precipitation), and (ii) for grid points that have non-zero precipitation in one or both fields, the data are highly skewed. One way to handle the skewness in the precipitation distribution at nonzero grid points is to try to remove it by transforming the fields (e.g., by taking the cubed roots of each field, $O^{1/3}$ and $F^{1/3}$). The zero inflation can be handled by applying a mixture of distributions with a normal df with a very small variance around zero for the zero-valued grid points (e.g., Gilleland et al., 2010b). For this setup, and using $J(\boldsymbol{W})$ in place of $\sigma_\Delta^{-1}(\boldsymbol{I}-\boldsymbol{C})$ for the prior covariance on control point movements, the objective function becomes

$$Q(\boldsymbol{p}^F) = \sum_{i=1}^{n} \left\{ \frac{\beta}{2\sigma_\varepsilon^2} \left(O^{1/3}(\boldsymbol{s}_i) - \widetilde{F}^{1/3}(\boldsymbol{s}_i) \right)^2 + \frac{1-\beta}{2\times 10^{-4}\sigma_\varepsilon^2} \left(O^{1/3}(\boldsymbol{s}_i) - \widetilde{F}^{1/3}(\boldsymbol{s}_i) \right)^2 \right\} +$$
$$\frac{1}{2\sigma_\Delta} \left((\boldsymbol{p}_x^F - \boldsymbol{p}_x^O)^T \mathcal{J}^{-1}(W_x)(\boldsymbol{p}_x^F - \boldsymbol{p}_x^O) + (\boldsymbol{p}_y^F - \boldsymbol{p}_y^O)^T \mathcal{J}^{-1}(W_y)(\boldsymbol{p}_y^F - \boldsymbol{p}_y^O) \right),$$

where $\mathcal{J}(\cdot)$ is defined in Eq (6) below.

In this case, a parameter β is introduced that takes values in $[0, 1]$ providing the mixture between the two errors (one being the error at non-zero grid points, and one approximating a Dirac δ at zero-valued points). It is an estimate of the percentage of zero-valued grid points in both fields, which is not especially useful and is assumed a nuisance parameter. The values of σ_ε and σ_Δ can be thought of as the variance of the errors and control point movements. However, in this context it is useful to choose their values *a priori* in order to obtain meaningful deformations. Therefore, they too are treated as nuisance parameters.

All of the above applies to a single *snapshot* of a verification set. A spatial displacement error may, or may not, be the result of a timing error. When applied only to a single time point, image warping gives useful information, but it may be possible to obtain an even more informative warp if temporal information could

also be considered. It turns out that it is relatively straightforward to extend the concept to further dimensions, at least mathematically, provided the resolution in the extra dimension(s) is comparable to the spatial dimension.

Here, the temporal dimension is treated. In particular, forecasts with the same initialization time with lead times at $t-1$, t and $t+1$ are utilized and compared to observation fields at time t. For example, if t represents the 24-h precipitation forecast, then the question concerns how well the 24-h lead time performs. The image warp can draw upon the 23-h lead time forecast as well as the 25-h lead time (it is assumed that 1-h differences are comparable in resolution to the spatial grid spacing). Of course, one could consider other lead times as appropriate.

For the space-time warp applied here, the Gaussian Markov Random Field approach is used. Control points (as all other points) are now three-dimensional, (x, y, t), and the objective function Q is a simple extension of (9). For example, letting $\boldsymbol{\Delta}_t = \boldsymbol{p}_t^F - \boldsymbol{p}_t^O$, the new term $[\boldsymbol{\Delta}_t^T(\boldsymbol{I}-\boldsymbol{C})\boldsymbol{\Delta}_t]/(2\sigma_t^2)$ is added to the end. The new penalty for undesirable deformations in the temporal dimension (σ_t) is taken here to be the same as σ_Δ, the penalty for undesirable deformations in space, which seems to be a reasonable strategy (further study is needed, however). In general, allowing σ_t to differ would remove the requirement that the scale be similar in time and space.

3 Some issues and details

Because the verification sets (i.e., the observed and forecast fields) have discrete representations (grid points), and the mapping (2) is continuous, care must be taken in determining values at grid points. That is, a map of a point (x, y) will in most cases not lie exactly in the middle of a grid point in the undeformed field. To determine what intensity value should be assigned to the point (x, y), some interpolation of the values in the surrounding grid points in F must be used. For example, one might use bilinear interpolation, however bicubic interpolation may prove useful because it will allow for analytic calculation of higher order derivatives for the optimization routines (e.g., Gilleland et al., 2010b).

Specifically, let r and s represent the fractional part of $W_x(x, y)$ and $W_y(x, y)$, respectively. That is,

$$r = W_x(x, y) - \lfloor W_x(x, y) \rfloor$$
$$s = W_y(x, y) - \lfloor W_y(x, y) \rfloor,$$

where $\lfloor x \rfloor$ is the greatest integer less than x. The value at a grid point in the deformed field, $\widetilde{F}$, is then interpolated by

$$\widetilde{F}(x, y) = \sum_l \sum_k b_l(r) b_k(s) F(\lfloor W_x(x, y) \rfloor + l, \lfloor W_y(x, y) \rfloor + k),$$

where $b_l(r)$ and $b_k(s)$, and their range, depend on the type of interpolation. For bilinear interpolation, the range for both is 0 to 1 (the center of a pixel/grid-point being at 1/2), and they are given by $b_0(t) = 1 - t$ and $b_1(t) = t$. For bicubic interpolation, the range for both is from -1 to 2, and the interpolation polynomials are given by

$$b_{-1}(t) = (2t^2 - t^3 - t)/2$$
$$b_0(t) = (3t^3 - 5t^2 + 2)/2$$
$$b_1(t) = (4t^2 - 3t^3 + t)/2$$
$$b_2(t) = ((t-1)t^2)/2.$$

Another issue that needs addressing concerns points that are mapped outside the field's domain (i.e., $(W_x(x,y), W_y(x,y)) \notin \mathcal{D}$). In such a case, those grid points require extrapolation, which can be done using any type of spatial prediction/interpolation approach (e.g., kriging). For some fields, such as precipitation, where there are many zero-valued grid points, it may suffice to simply set those values outside the domain to zero. Other approaches also exist, however. Regardless, if there is a lot of important activity near the edges of the domain, the deformation approach may be questionable. This is true of most spatial forecast verification approaches, and typically, the forecast domain centers on the area of interest so that this should not be a major concern.

For determining derivatives, the fields we use may not be "smooth" enough, so some smoothing may be necessary for the optimization routines.

4 Analyzing the results

Gilleland et al. (2010b) propose a method for summarizing the results of image warping for the purpose of ranking multiple forecasts. They applied this technique to a host of test cases, and found that the summaries provided useful information. The main drawback to the summary is in its complexity. Heuristically, the best forecast is the one that requires the least amount of movement (better termed *bending energy*) and yields the highest improvement in the traditional grid-point to grid-point statistics. Unfortunately, this is complicated by the fact that a perfect forecast would require no deformation, and subsequently would not have any improvement in the statistic. Further, a very poor forecast might have a very large reduction in error. Therefore, some care needs to be taken in summarizing forecast performance. This is the reason for the complexity in the IWS score proposed by Gilleland et al. (2010b), requiring some user-chosen weights, but fortunately results are relatively insensitive to these choices.

Summary statistics are required for many applications, but the real power of most spatial verification techniques lies in their ability to provide useful diagnostic information. Often, this cannot be done for large numbers of cases (e.g., in an

Figure 1: WRF ARW 24-h accumulated precipitation (forecast) and Stage II analysis (observation) valid 1 June 2005 (one of the nine real ICP test cases). Grid-point to grid-point verification results (RMSE, MAE and ME) before and after image warping are shown.

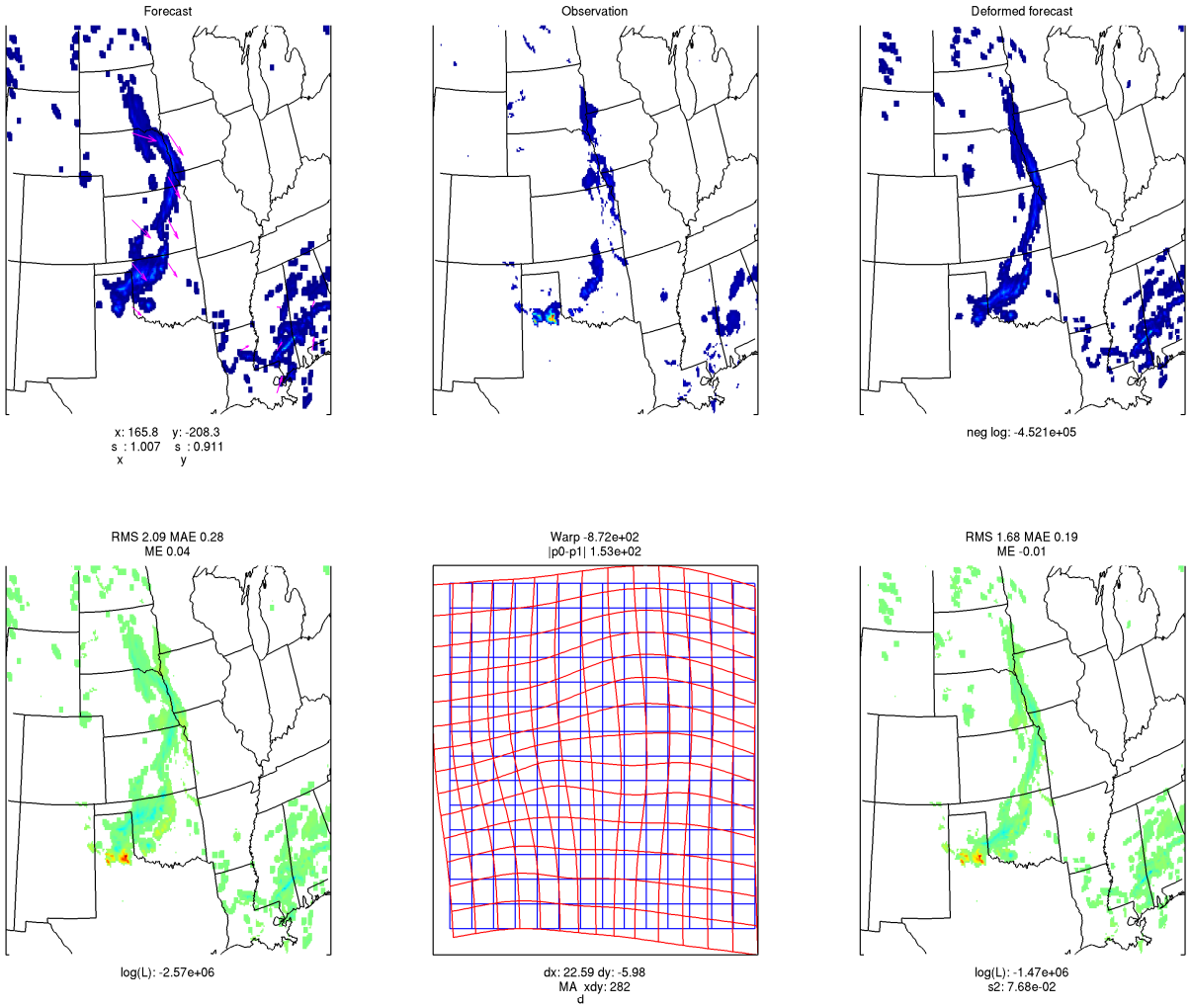
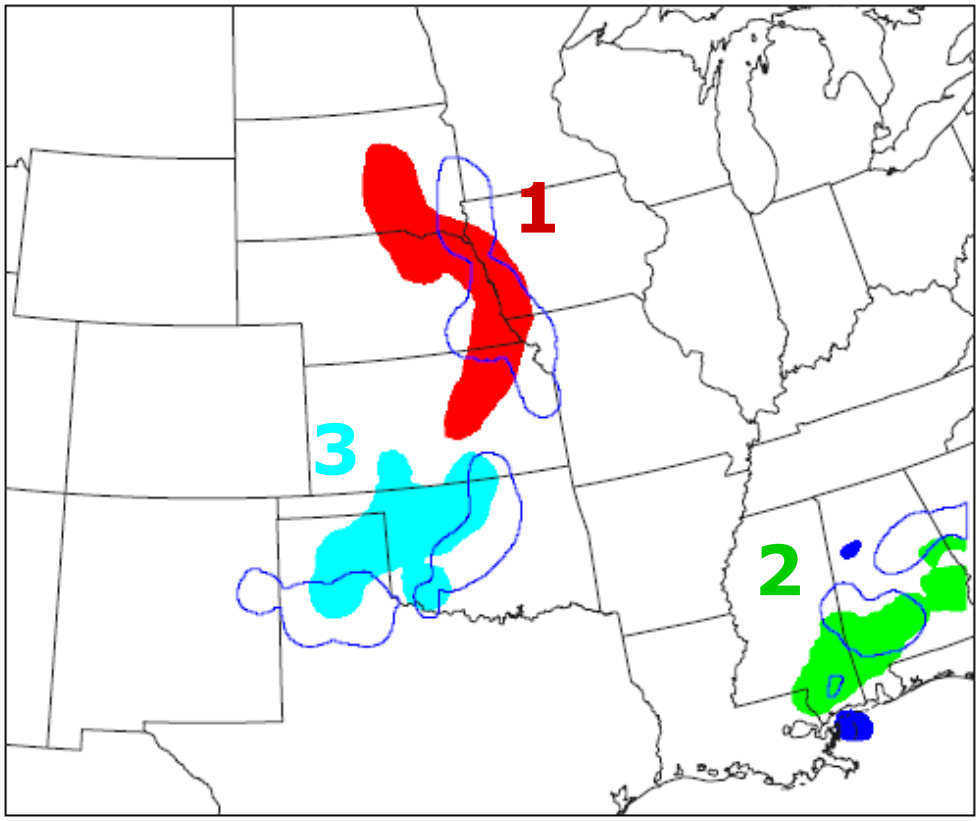


Figure 2: Objects defined by MODE for WRF ARW 24-h accumulated precipitation (forecast, solids) and Stage II analysis (observation, outlines) valid 1 June 2005 (one of the nine real ICP test cases).



operational setting), but is still useful for many purposes (e.g., in a research setting). Circle histograms showing the distribution of the amount of displacement in different directions is potentially a very useful summary tool (not done here). This would give the user a feel for whether a model makes systematic errors in a particular direction. Applying such histograms to smaller regions within a larger region is also probably a good idea as errors in certain directions may vary regionally (e.g., Fig. 1).

4.1 Spatial Image Warping

In both Figs. 1 and 2, it can be seen that the forecast was spatially displaced by being generally too far north in the central part of the region, but is mostly too far south in the southeastern region. Image warping, therefore moves the central activity predominantly south, while it moves the southeastern forecasted precipitation

northwards. The features-based procedure known as MODE (Davis et al., 2009) also picks up on these discrepancies; recall that features-based procedures handle features individually so that the forecast features identified in Fig. 2 are labeled 1 (red), 2 (green) and 3 (light blue). The dark blue features were not matched to anything in the analysis field, and the blue outlines identify the associated analysis features.

Fig. 1 shows different ways of visualizing the image warp procedure. In the upper left most panel, the original (undeformed) forecast field is shown with vectors super-imposed to show how the image warp is deforming the field. The resulting deformed forecast is shown in the upper right corner, and it is easy to see that it matches better with the analysis field (top middle) than the undeformed forecast. Grid-point to grid-point differences are shown between the undeformed forecast field and analysis (lower left panel) and between the deformed forecast and analysis (lower right panel). While errors still exist, they are fewer for the deformed forecast. Summary statistics also show this. In particular, the RMSE is 2.09 originally, and this is reduced to 1.68 after deformation. Therefore, 1.68 represents the RMSE after accounting for displacement errors. The middle panel of the bottom row demonstrates what the deformation does to the grid itself. Recall that image warping is akin to moving, rotating stretching, and "crinkling" the grid (i.e., intensities are not involved). While possible to interpret this panel in terms of how the deformation is made, it is less straightforward, but it does show visually how much and how complex the deformation is. It also shows that the deformation is physically meaningful (e.g., it does not fold over onto itself). This procedure was applied to all nine of the real ICP test cases (see e.g., Ahijevych et al., 2009). See Gilleland et al. (2010b) for the results, and results from a larger set of 32 cases that were also analyzed with MODE (Davis et al., 2009).

Fig. 3 demonstrates the image warp deformations for the five geometric cases of the ICP shown in Fig. 4. In each case, the deformed image is almost exactly identical to the observed field (negligible grid-point to grid-point RMSE's). this demonstrates that even when the assumption of normality for the error term in (1) is wildly violated, as it is here, the resulting deformations are still valid. As usual, it is when inferences are to be made concerning the deformation that the distributional assumption becomes important. This is especially true when using image warping to forecast ahead in time (e.g., as in Åberg et al., 2005). In the context of forecast verification, however, this is less of an issue. Although it is important to characterize the uncertainty in the verification results, confidence intervals on the deformations may be very difficult to interpret, and are certainly difficult to display. While the procedure is set up in a statistically elegant fashion that enables a straightforward framework with which to construct such intervals, more work needs to be done to determine a useful way to convey them, and to

Figure 3: Image warping applied to the five ICP geometric cases (see Fig. 4), and one additional case (lower right panel).

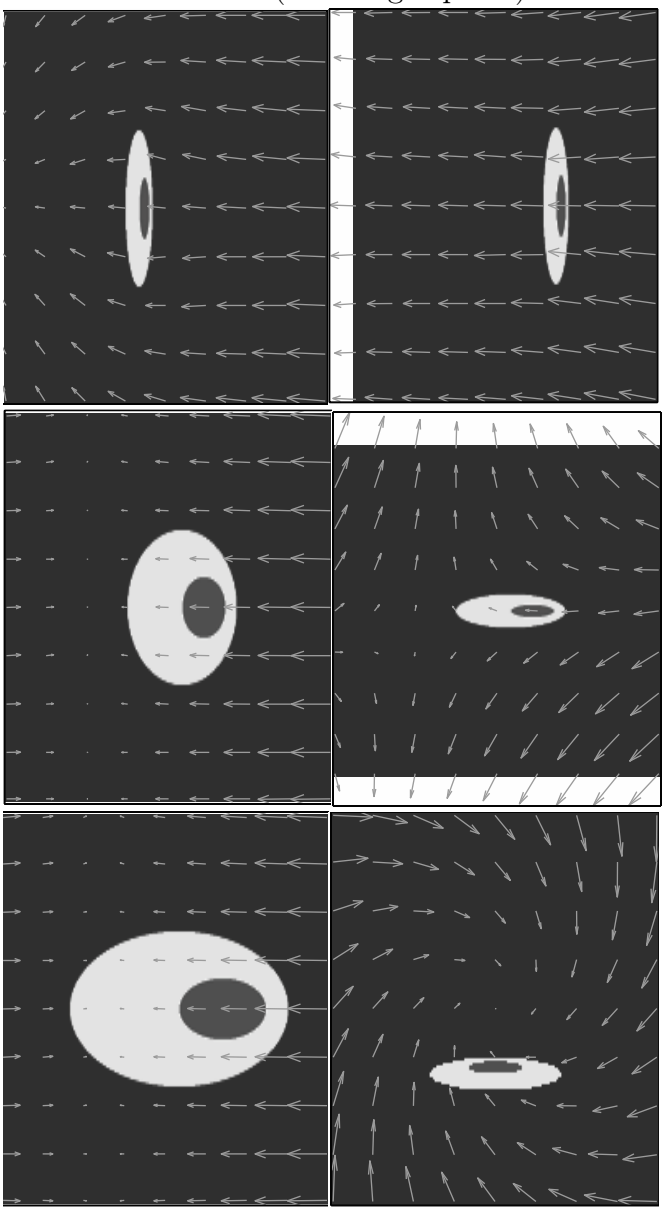
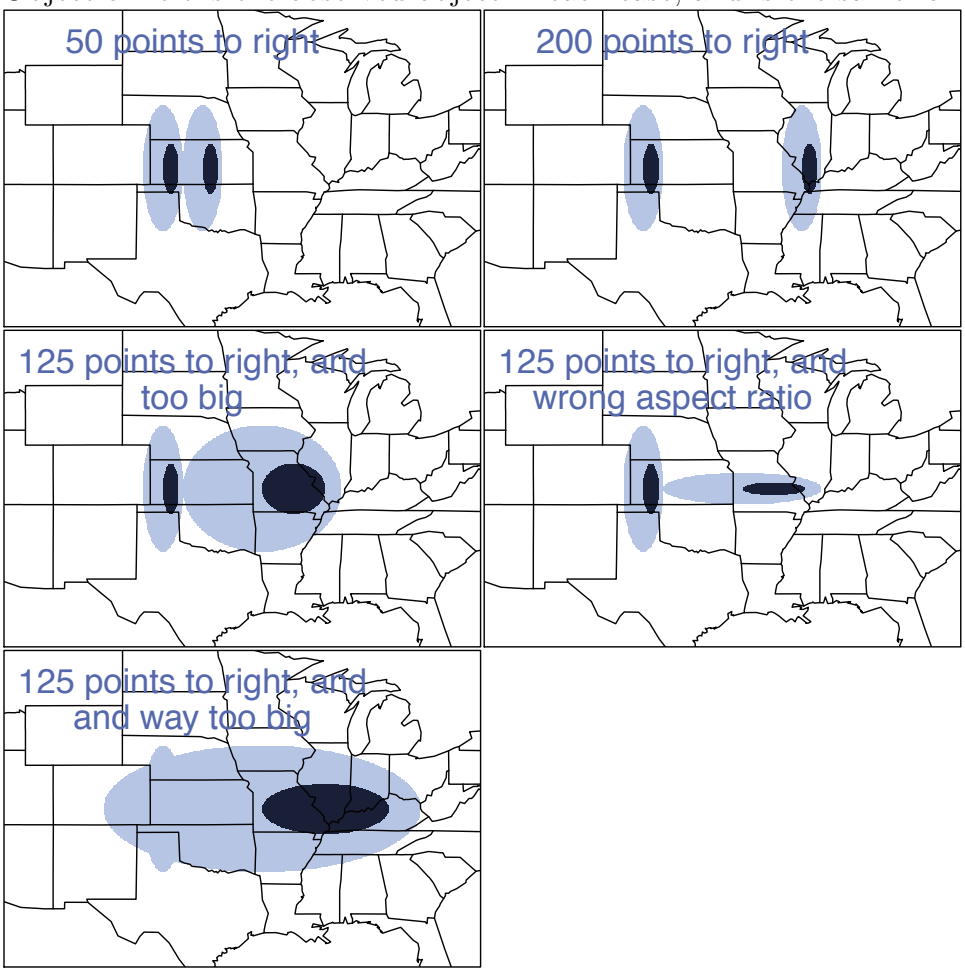


Figure 4: The five ICP geometric cases along with their respective known errors. Object on left is the observed object in each case, and is the same for each.



propagate them to summary statistics, such as the IWS (Gilleland et al., 2010b).

One issue that arises concerns whether or not the image warp procedure produces the "correct" deformations when they are known. The geometric cases demonstrate quite clearly that it does. In fact, it was thought to include a *rotation* error in the cases. However, the resulting case was actually an aspect ratio error, and the image warp procedure first brought this to light; showing that it can aid the subjective observer. To show that the technique could decipher rotation from aspect ratio errors, image warping was applied to a true rotation error (Fig. 3, lower right panel), and it clearly could. In practice, this is not of great importance, but it demonstrates the accuracy of the technique. Further, with careful programming, image warping can be done efficiently enough as to be performed operationally (each case shown here was carried out on an ordinary laptop at about five minutes per case).

4.2 Space-time Image Warping

Using a larger set of data from the same National Severe Storms Lab (NSSL)/Storm Prediction Center (SPC) 2005 Spring Program that is better resolved in time (hourly) than the ICP real test cases are used to obtain an initial look at adding the temporal component to the spatial image warp (also, instead of 24-h accumulations, only 1-h accumulations are used here). In order to better focus on the ability of the image warp to adequately deform through both space and time, only a relatively small region is chosen for analysis (Fig. 5). Further, only NCAR and NCEP WRF $\approx$ 4-km models are studied. Results for 26 April 2005 at the 20th hour are shown here, though other times were also studied with similar results (in terms of the image warping procedure). That is, the analysis is for 26 April 2005 at 20h UTC, and time t corresponds to the forecast valid on this date at the 20th hour (UTC), $t-1$ is the forecast valid at the 19th hour, and $t+1$ at the 21st hour. All forecasts were initialized at the same time (0 UTC).

Figs. 6 and 7 show image warping results for the NCAR and NCEP, resp., models for 26 April 2005. The valid time is for $t = 20$ hours UTC on this date. This case was chosen in part because both models had deformations that moved in both space and time (many cases did not require movement through time). It is clearly difficult to decipher subjectively whether points should be moved across a temporal coordinate versus through a spatial one, but careful study suggests that the resulting deformation is reasonable. The NCAR model for time t shows heavy precipitation in the northern part of Michigan, where the analysis has it further south, closer to the center of the state. Its forecast for $t + 1$, on the other hand, has some heavy activity closer to where it occurs in the analysis, but too far south. The image warp subsequently tries to move the $t + 1$ activity to the north (and back in time), while also pulling some values from $t - 1$, although it is

Figure 5: Domain used for the space-time analysis.

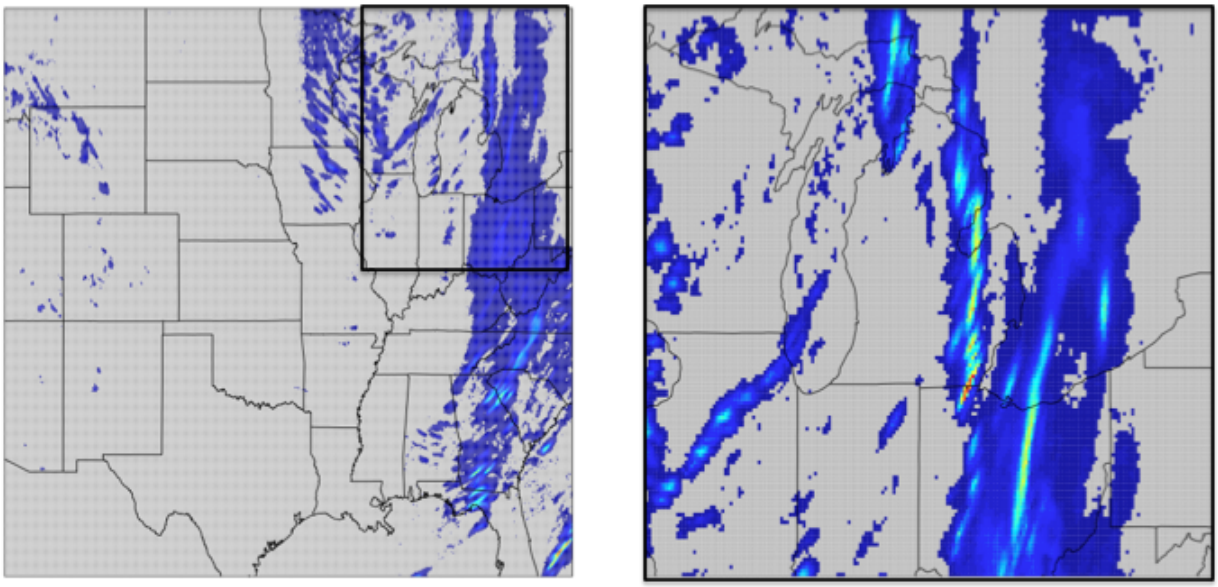


Figure 6: Space-time image warp applied to NCAR's 4-km WRF 1-h accumulated precipitation (mm) and Stage II analysis at this valid time.

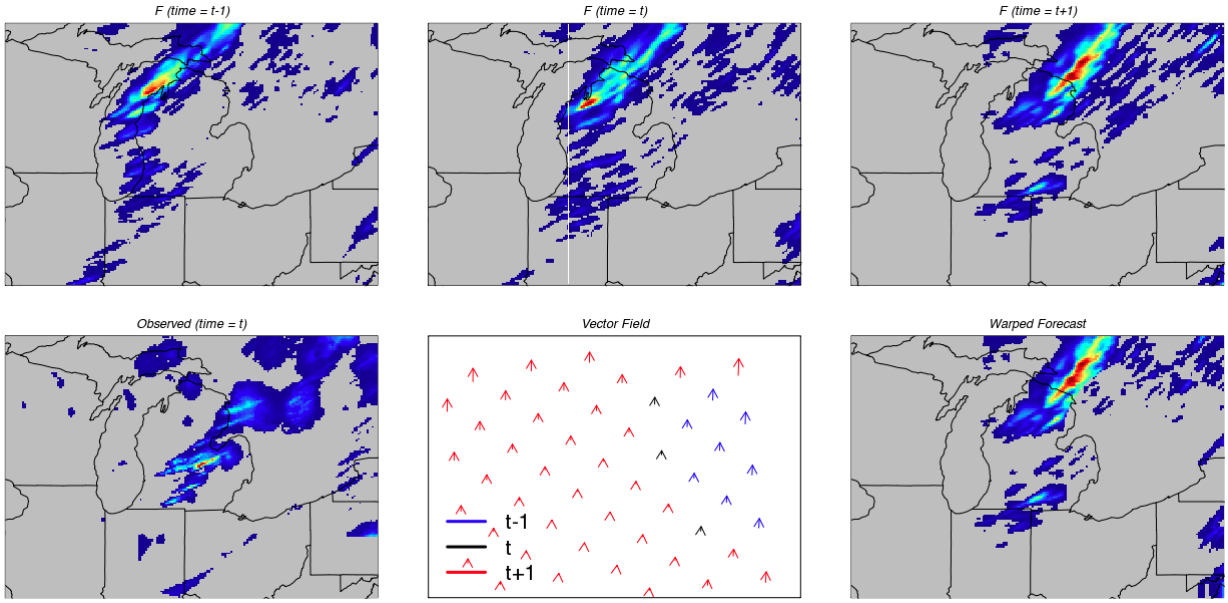


Figure 7: Space-time image warp applied to NCEP's 4-km WRF 1-h accumulated precipitation (mm) and Stage II analysis at this valid time.

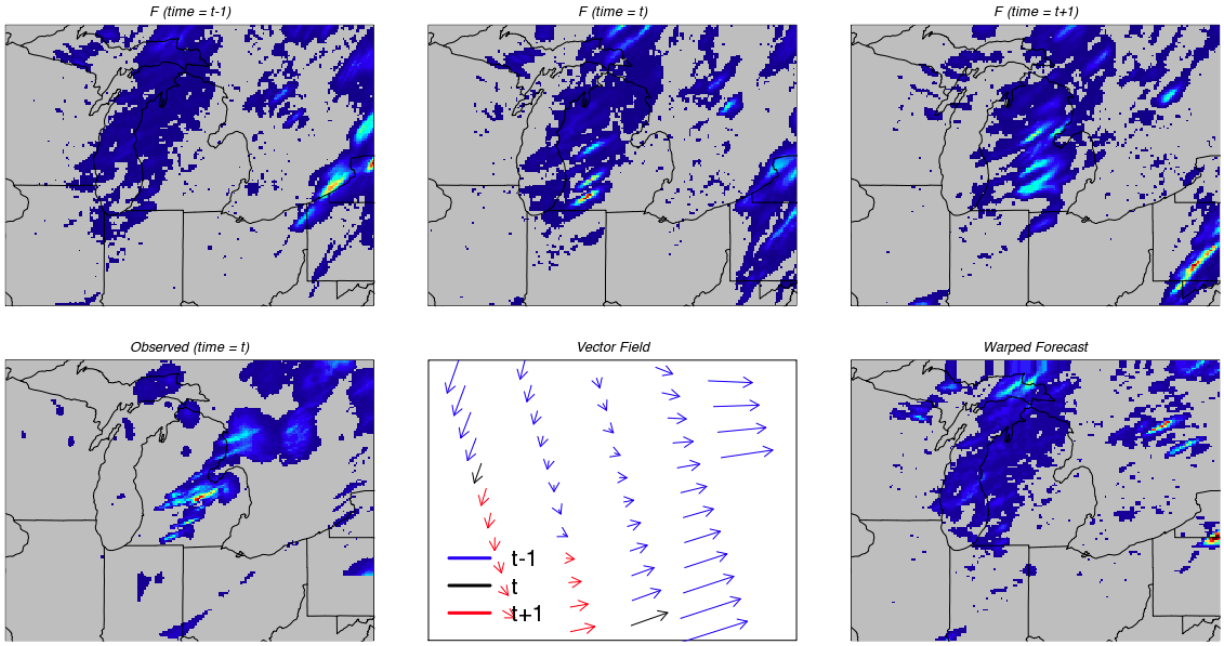


Table 1: Space-Time Image Warp verification results for 26 April 2005 comparison of NCAR and NCEP WRF models.

	Loss function value		RMSE		reduction in RMSE
	no warp	warp	no warp	warp	
NCAR	8010.1	2657.2	2.67	0.73	56.5%
NCEP	3361.6	1243.1	2.24	1.51	32.4%

Table 2: Space-Time Image Warp verification results for 20 May 2005 comparison of NCAR and NCEP WRF models.

	Loss function value		RMSE		reduction in RMSE
	no warp	warp	no warp	warp	
NCAR	28.51	8.03	0.04	0.009	74.4%
NCEP	15.93	6.77	0.02	0.009	57.4%

less clear that this is necessary. For the spatial case, Gilleland et al. (2010b) found that the image warp can be sensitive to a lot of scatter, which is easily remedied by setting a threshold to filter out activity that may not be interesting. Perhaps a more interpretable space-time warp could also be obtained by such a scheme. The NCEP model also does not appear to capture the heavy activity in the center of the state, but it does have some higher values in that area. Overall, however, it would seem that the NCEP model greatly over forecasts the spatial extent of lower intensity precipitation. Again, setting a threshold should help to resolve this issue.

To better determine how well the image warp performed through both space and time, it is useful to observe some summary statistics. Table 1 shows the before- and after- deformation values for the loss function Q, and RMSE. It also shows the percent reduction in RMSE betemed by the deformation. The NCEP model fairs better than the NCAR model in terms of the initial RMSE and in terms of the before and after value of the loss function, but this may be a result of the frequency bias of small values by the NCEP model. This can be seen by the fact that less energy was spent to deform the NCAR model, while a higher percent reduction in RMSE is obtained along with a lower resulting RMSE. That is, for this case, values of the NCAR model require less movement, and result in a deformed field that is *closer* to the analysis field. Similar results are obtained for another case (valid on 20 May), but where this time there is considerably less rain activity in the region. Subsequently, initial traditional score values (RMSE) are lower, and both drop to about 0.009. Considerably more morphing is required for the NCEP configuration than NCAR's, however, and again the percent reduction in the RMSE is much greater for NCAR's configuration ($\approx 15\%$ difference).

5 Summary and Conclusions

High-resolution forecast verification applied to an observed field on the same grid is a challenging new area of research. Many procedures have been proposed in the last decade to better ascertain forecast performance in this context. An international inter-comparison project, the ICP, has been established to better understand the many new methods, and how they inform about forecast quality. Initial results from the ICP pertaining to QPF fields have been presented in a special collection of papers for the journal *Weather and Forecasting*. As part of these initial findings, most methods can be classified as belonging (if only loosely for some methods) to one of four categories. Two of these categories are filter based, and the other two are displacement based (Gilleland et al., 2009, 2010a). This paper discusses a field deformation approach that belongs to the latter category, called image warping.

Gilleland et al. (2010b) showed that image warping works well for the spatial snapshots of a gridded verification set. They also introduced a summary statistic that allows for comparison of multiple forecasts using the image warp results. This paper gives some more technical details about the procedure utilized therein.

Beyond the spatial snapshot case, an initial look at incorporating both space and time is given here. Initial results applied to a limited set of cases show that the approach has merit, and further study is recommended to determine both how best to carry out the technique, as well as how to interpret the results.

In particular, careful study of the parameter values for the space-time case needs to be conducted. For example, in the present study, the temporal penalty is held to be the same as the spatial one. Perhaps this is not the optimal choice, however. To some extent, it will depend on the specific model or user need. However, it would be useful to have some guidance about what values are most appropriate. Similarly, the present study assumes that the temporal spacing is comparable to the spatial resolution. That is, one hour is assumed to be comparable to about 4 km of space. This assumption should be checked. For both of these issues, it may suffice to look at simple geometric cases, similar to what has been done in the ICP. For example, it may be useful to look at an observed circle and warp against a circle in the same spot but at different times (with no activity in the other fields). What happens when $t-1$ and $t+1$ have the circle, but t has nothing?

It is also important to investigate whether the image warp pulls forecast rain from time when a human observer would prefer to pull it from space. If this is the case, it can most likely be remedied by changing the penalties for space and/or time, but it is important to obtain guidance about proper values to use.

Are the methods for comparing multiple forecasts given in Gilleland et al. (2010b) naturally extendable to the space-time case? If not, is there a way to summarize the results both for comparing multiple forecasts and for summarizing a single forecast's performance?

More technical issues that need to be addressed include the aforementioned question about the calculation and communication of uncertainty. Of course, this needs to be addressed for the simpler case of space only. Additionally, it has been assumed that space and time are independent in regards of the penalty term of the loss function. This may be perfectly reasonable from a practical point of view, but it should be investigated from a statistical perspective.

It was also mentioned that one could use other time points besides $t-1$, t and $t+1$. In fact, it may depend on the forecast, and an initial subjective opinion would likely determine the best set to include. Inclusion of $t \pm k$ times, however, will need to be chosen carefully, and some of the concerns stated above may need to be re-checked.

Overall, the image warp technique is a valid, useful, and highly accurate procedure for analyzing forecast performance. It not only can yield summary measures, but can also provide a wealth of diagnostic information. An initial study of including time and space has shown that this is a potentially very helpful addition to the framework, and that it is possible to do.

Acknowledgments

We thank Michael Baldwin (Purdue) for making the real ICP data cases available for us, and David Ahijevych (NCAR) for the geometric cases. Support for this manuscript was provided by the Air Force Weather Agency (AFWA, http://www.afweather.af.mil/), the Developmental Testbed Center (DTC, http://www.dtcenter.org/index.php) and the National Center for Atmospheric Research (NCAR). Participation by Johan Lindström and Finn Lindgren was made possible by the Swedish Foundation for International Cooperation in Research and Higher Education (STINT) Grant IG2005-2047. Participation of all other co-authors was made possible by the Statistical and Applied Mathematical Sciences Institute (SAMSI) and the Center for Research in Scientific Computation (http://www.ncsu.edu/crsc/) of North Carolina State University, Raleigh by way of the Industrial Mathematical and Statistical Modeling Workshop (IMSM, http://www.ncsu.edu/crsc/events/imsm10/). NCAR and SAMSI are sponsored by the National Science Foundation. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Åberg, S., F. Lindgren, A. Malmberg, J. Holst, and U. Holst, 2005: An image

warping approach to spatio-temporal modelling. *Environmetrics*, **16**, 833–848.

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

Alexander, G., J. Weinman, V. Karyampudi, W. Olson, and L. A.C.L., 1999: The effect of assimilating rain rates derived from satellites and lightning on forecasts on the 1993 superstorm. *Mon. Wea. Rev.*, **127**, 1433–1457.

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

Casati, B., 2009: New developments of the intensity-scale technique within the spatial verification methods inter-comparison project. *Submitted to Wea. Forecasting*.

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

Davis, C., B. Brown, and R. Bullock, 2006a: Object-based verification of precipitation forecasts, part i: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev.*, **134**, 1772–1784.

— 2006b: Object-based verification of precipitation forecasts, part ii: Application to convective rain systems. *Mon. Wea. Rev.*, **134**, 1785–1795.

Davis, C., B. Brown, R. Bullock, and J. Halley Gotway, 2009: The method for object-based diagnostic evaluation (MODE) applied to numerical forecasts from the 2005 NSSL/SPC spring program. *Wea. Forecasting*, **24**, 1252–1267.

Ebert, E., 2008: Fuzzy verification of high resolution gridded forecasts: A review and proposed framework. *Meteorol. Appl.*, **15**, 51–64, doi:DOI: 10.1002/met.25.

Ebert, E. and W. J. Gallus, 2009: Toward better understanding of the Contiguous Rain Area (CRA) method for spatial forecast verification. *Wea. Forecasting*, **24**, 1401–1415.

Ebert, E. and J. McBride, 2000: Verification of precipitation in weather systems: determination of systematic errors. *J. Hydrology*, **239**, 179–202.

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

Gilleland, E., D. Ahijevych, B. Brown, and E. Ebert, 2010a: Verifying forecasts spatially. *Bull. Amer. Meteorol. Soc.*, October, 1365–1373.

Gilleland, E., T. Lee, J. Halley-Gotway, R. Bullock, and B. Brown, 2008: Computationally efficient spatial forecast verification using baddeley's δ image metric. *Mon. Wea. Rev.*, **136**, 1747–1757.

Gilleland, E., J. Lindström, and F. Lindgren, 2010b: Analyzing the image warp forecast verification method on precipitation fields from the ICP. *Wea. Forecasting*, **25**, 1249–1262.

Glasbey, C. and K. Mardia, 1998: A review of image warping methods. *Journal of Applied Statistics*, **25**, 155–171.

Grams, J., W. Gallus Jr., S. Koch, L. Wharton, A. Loughe, and E. Ebert, 2006: The use of a modified ebert-mcbride technique to evaluate mesoscale model qpf as a function of convective system morphology during ihop 2002. *Wea. Forecasting*, **21**, 288–306.

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

Hoffman, R., Z. Liu, J. Louis, and C. Grassotti, 1995: Distortion representation of forecast errors. *Mon. Wea. Rev.*, **123**, 2758–2770.

Keil, C. and G. Craig, 2007: A displacement-based error measure applied in a regional ensemble forecasting system. *Mon. Wea. Rev.*, **135**, 3248–3259.

— 2009: A displacement-based error measure applied in a regional ensemble forecasting system. *Wea. Forecasting*, **24**, 1297–1308.

Marzban, C. and S. Sandgathe, 2006a: Cluster analysis for ob ject-oriented verification of fields: A variation. *Submitted to Mon. Wea. Rev*.

— 2006b: Cluster analysis for verification of precipitation fields. *Wea. Forecasting*, **21**, 824–838.

— 2009: Verification with variograms. *Accepted to Wea. Forecasting*.

Mass, C., D. Ovens, K. Westrick, and B. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? the results of two years of real-time numerical weather prediction over the pacific northwest. *Bull. Amer. Meteorol. Soc.*, March, 407–430.

Nachamkin, J., 2004: Mesoscale verification using meteorological composites. *Mon. Wea. Rev.*, **132**, 941–955.

— 2009: Application of the composite method to the spatial forecast verification methods intercomparison dataset. *Wea. Forecasting*, **24**, 1390–1400.

Nehrkorn, T., R. Hoffman, C. Grassotti, and J.-F. Louis, 2003: Feature calibration and alignment to represent model forecast errors: Empirical regularization. *Q.J.R. Meteorol. Soc.*, **129**, 195–218.

Venugopal, V., S. Basu, and E. Foufoula-Georgiou, 2005: A new metric for comparing precipitation patterns with an application to ensemble forecasts. *J. Geophys. Res.*, **110**, D8, D08111 10.1029/2004JD005395.

Wernli, H., C. Hofmann, and M. Zimmer, 2009: Spatial forecast verification methods inter-comparison project – application of the SAL technique. *Wea. Forecasting*, **24**, 1472–1484.

Wernli, H., M. Paulat, M. Hagen, and C. Frei, 2008: SAL–a novel quality measure for the verification of quantitative precipitation forecasts. *Mon. Wea. Rev.*, **136**, 4470–4487.