A Gaussian Mixture Model Approach to Forecast Verification

Valliappa Lakshmanan[1,2]*, John S. Kain[2]

*Corresponding author: V Lakshmanan, 120 David L. Boren Blvd, Norman OK 73072; lakshman@ou.edu

[1]Cooperative Institutute of Mesoscale Meteorological Studies, University of Oklahoma; [2]National Oceanic and Atmospheric Administration / National Severe Storms Laboratory

ABSTRACT

Verification methods for high-resolution forecasts have been based either on filtering or on objects created by thresholding the images. The filtering methods do not easily permit the use of deformation while identifying objects based on thresholds can be problematic. In this paper, we introduce a new approach in which the observed and forecast fields are broken down into a mixture of Gaussians, and the parameters of the Gaussian Mixture Model fit are examined to identify translation, rotation and scaling errors. We discuss the advantages of this method in terms of the traditional filtering or object-based methods and interpret resulting scores on a standard verification dataset.

```
@Article{gmmverif,
  author =       {Lakshmanan, V. and Kain, J.},
  title =        {A {G}aussian Mixture Model Approach to Forecast
                  Verification},
  journal =      {Weather and Forecasting},
  year =         2010,
  volume =       25,
  number =       3,
  pages =        {908-920}
}
```

1. Introduction

Intuitively, approximating a gridded field by a Gaussian Mixture Model (GMM) may be thought of as the process of finding an optimal way to place Gaussian functions at various points in the image such that the sum of these Gaussians mimics the input gridded field. As shown in Figure 1, the larger the number of Gaussian components in the mixture model, the more closely the image recreated using just the Gaussian components resembles the original image.

Given the GMM that approximates two images (the forecast and observed), we show in Section 3 that it is possible to analyze the parameters of the component Gaussians to infer translation, rotation and scaling transformations.

a. Relationship to verification approaches

The new methods of verifying model forecasts that have been proposed can be categorized into (a) filtering-based methods that operate on the neighborhood of pixels or on the basis of decomposition and (b) displacement-methods that rely either on features or on field deformation (Gilleland et al. 2009). Here, we propose a method of verification that does not quite fall into any of these categories.

Our proposed method incorporates level of detail, like the filtering methods, in that the approximation can be made as exact as desired by increasing the number of Gaussian components allowed in the mixture. The most exact representation would be a mixture of Gaussians of zero variance and a component centered at every grid point. However, our proposed method operates neither on the neighborhood of pixels nor on the basis of wavelet-

like decompositions.

We propose analyzing the entire image (like field deformation), but only to find a parametric approximation to the image. Field deformation approaches such as those of Alexander et al. (1999); Keil and Craig (2007) employ non-parametric optical flow approaches. In our approach, the parameters of the approximation are compared between the forecast and observed fields to obtain insight into the transformations (translation, rotation and scaling) that would make the fields most like each other.

In the use of transformations, the method of this paper resembles the feature-based approaches of Davis et al. (2006) but without the dependence on thresholds (either in intensity or in size) to categorize "objects". Therefore, our approach is not quite "object-based". It could, however, be considered feature-based if one were to extend the definition of "feature" to include the Gaussian components that form the mixture.

It should be noted that the GMM approach does require a threshold – only pixels with intensity above that threshold will be considered in the GMM fit. For the precipitation forecast fields of Figure 1, only pixels with rainfall amounts greater than 6.6mm, corresponding to the top 10% of the pixel values in the image, were used to fit the GMM whereas for the synthetic fields of Figure 2, only pixels with non-zero values were fit by the GMM. The difference between the GMM approach of this paper and the object-based approach of Davis et al. (2006) is that in the GMM approach, this threshold does not determine what the "objects" are. Thus, as shown in Figure 1, one could have either 3 features or 50 by choosing to fit all the pixels in the image above the 6.6mm threshold to either a GMM with 3 components or to one with 50 components.

b. Advantages of the GMM approach

There are several advantages to fitting an image with a GMM and using the fitted GMM to carry out forecast verification:

i. There is no need to be concerned with splits or merges – if two contiguous regions are better treated as a single region, then they will be approximated by a single Gaussian. Conversely, a single, contiguous region may be broken up into multiple Gaussians if needed for an optimal fit and if there are enough GMM components.

ii. The Gaussian is a parametric function. Thus, the GMM affords a highly compressed view of the information in the data that is especially useful for comparing two images for correspondence.

iii. The number of Gaussians used is a good measure of the level of detail at which the image is being represented. For the verification problem, by changing the number of Gaussians allowed in the mixture model, one can control the scale at which comparisons are carried out.

iv. Transformations of Gaussians correspond to easily identifiable changes in their parameters. Translation of objects corresponds to a change in the center point of the Gaussian. Scaling (corresponding objects being smaller or larger in one of the fields) can be inferred by changes in the variance of the Gaussian. Rotation of objects can be inferred by changes in the ratio of the variance of the Gaussian in east-west and north-south directions. Changes in the amplitude of the Gaussian correspond to changes in intensity.

The natural incorporation of level of detail is an important characteristic of filtering-based methods. The natural incorporation of transformation is a key advantage of object-based verification methods, especially because the detection of transformation permits verification methods to avoid the "double penalty" (Gilleland et al. 2009) problem. Thus, a GMM provides the advantages of both of these methods in a simple, mathematically elegant framework that is also quite easy to implement.

The method by which a GMM is fit to forecast and observed fields is described in Section 2. We present the results of comparing the GMM on fake geometric and perturbed cases drawn from Ahijevych et al. (2009) and Kain et al. (2008) and make suggestions for further work in Section 3.

2. Fitting a Gaussian Mixture Model

Fitting a GMM to an image for the purposes of forecast verification consists of the following steps:

i. Initialize the GMM (Section 2c).

ii. Carry out Expectation-Minimization (EM) algorithm to iteratively "tune" the GMM (Section 2b).

iii. Store the parameters of each Gaussian component of the GMM (Section 2d).

iv. Compute translation, rotation and scaling errors from the GMM parameters corresponding to the fits of the forecast and observed images (Section 2e).

The strategy followed for initializing the GMM will be much more clear if it is preceded by a mathematical description of the GMM and of the E-M algorithm. Hence, we define a GMM in Section 2a and explain the EM algorithm that is used to fit the image to the GMM in Section 2b before delving into the initialization strategy in Section 2c and listing the parameters to be stored in Section 2d. Error metrics are defined in Section 2e and these are used to determine the corresponding Gaussians in Section 2f.

a. The Gaussian Mixture Model (GMM)

The GMM is defined as a weighted sum of K two-dimensional Gaussians:

$$G(x,y) = \sum_{k=1}^{K} \pi_k f_k(x,y) \tag{1}$$

where the amplitudes π_k are usually chosen so that they sum to 1. Each of the two-dimensional Gaussians, $f_k(x,y)$ is defined given the parameters μ_{x_k}, μ_{y_k} and Σ_{xy_k} as (dropping the subscript k for convenience):

$$f(x,y) = \frac{1}{2\pi\sqrt{|\Sigma_{xy}|}} e^{-((x-\mu_x)(y-\mu_y))\Sigma_{xy}^{-1}((x-\mu_x)(y-\mu_y))^T/2} \tag{2}$$

μ_x, μ_y are the center of the Gaussian and Σ_{xy} the variance of the Gaussian i.e. Σ_{xy} is a matrix whose components are:

$$\begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix} \tag{3}$$

where σ_x is the standard deviation in the x direction and σ_{xy} the covariance of x and y. $|\Sigma_{xy}|$ is the determinant of the Σ_{xy} matrix. The scaling factor of the individual Gaussians $(1/(2\pi\sqrt{|\Sigma|}))$ has been chosen so that the Gaussians sum to 1 over all x, y. If the π_ks are

chosen to sum to 1, then the GMM also sums to 1 over the entire image. This allows a probabilistic formulation that will be taken advantage of shortly.

b. The Expectation-Minimization (EM) method

Given a set of points x_i, y_i, it is possible to fit these points to a GMM, $G(x, y)$, by following an iterative method known as the expectation-minimization (EM) method. The proof that this hill-climbing method works is available in many texts (e.g: Hand et al. (2001) pages 260-263), so we'll limit ourselves to describing the actual technique as it applies to the problem of fitting a GMM to the set of points.

Assume that an initial choice of parameters $\mu_{x_k}, \mu_{y_k}, \Sigma_{xy_k}$ exists for each of the K components. Because the scaling factors have been chosen to add up to one, the probability (or *likelihood*) that the point x_i, y_i is covered by the GMM given the set of parameters is given by:

$$P(x_i, y_i|\theta) = \sum_{k=1}^{K} \pi_k f_k(x_i, y_i|\mu_{x_k}, \mu_{y_k}, \Sigma_{xy_k}) \tag{4}$$

where θ is used as short-hand for all the parameters of all the K components.

The first step, known as the expectation-step or E-step, is to compute the likelihood of this given set of parameters. The probability that the pixel x_i, y_i arose from the kth Gaussian component is given by:

$$P(k|x_i, y_i, \theta) = \frac{\pi_k f_k(x_i, y_i|\mu_{x_k}, \mu_{y_k}, \Sigma_{xy_k})}{P(x_i, y_i|\theta)} \tag{5}$$

The second step, known as the minimization-step or M-step, is to update the parameters of all the K components based on the above likelihood calculations. To obtain the $\mu_x, \mu_y, \Sigma_{xy}$

of the kth component, the points x_i, y_i are weighted by $P_k(x_i, y_i)$ before the appropriate statistics are computed. For example,

$$\mu_x = E(x) = \frac{\sum_{i=1}^{N}(P_k(x_i, y_i)x_i)}{\sum_{i=1}^{N} P_k(x_i, y_i)} \tag{6}$$

Similarly, μ_y is computed as $E(y)$ and Σ_{xy} is computed as:

$$\begin{pmatrix} E((x-\mu_x)^2) & E((x-\mu_x)(y-\mu_y)) \\ E((x-\mu_x)(y-\mu_y)) & E((y-\mu_y)^2) \end{pmatrix} \tag{7}$$

Finally,the amplitude π_k is computed as:

$$\pi_k = \frac{1}{N}\sum_{i=1}^{N} P_k(x_i, y_i) \tag{8}$$

With the updated parameters, the E-step is carried out, a new set of likelihoods computed, used to weight the points in the next M-step, and so on until convergence is reached. The convergence is tested on the total likelihood of all the points at end of each M-step as follows.

Recall that the probability that the point x_i, y_i is covered by the GMM given the set of parameters is given by $P(x_i, y_i|\theta)$. From this, the probability that all the given points are covered by the GMM is given by the product of $P(x_i, y_i; \theta)$ over all the points. To avoid numerical instability errors when multiplying so many small numbers, the log of this likelihood is computed instead:

$$l(\theta) = \sum_{i=1}^{N} log(P(x_i, y_i)) \tag{9}$$

When the improvement in $l(\theta)$ falls below some tolerance, the iterative E-M process can be stopped. We stopped the E-M process when the improvement fell below 1% and found that convergence happens in 5 to 10 iterations.

The entire GMM fitting process is computationally very cheap. Each iteration of this process consists simply of computing weights by summing up previously computed values (Eq. 4,5) and then computing weighted averages (Eq. 6-8). We found that computing a 50-component GMM fit on a 500x600 image took just 0.05 seconds on a 1 GHz processor.

c. Initialization of the GMM

Recall that the E-step requires a set of components, and the weights computed at the end of the E-step are required to create a set of components in the M-step. Thus, the EM process has to be bootstrapped with some initial guess at a GMM. Then, the EM process will start at that point and slowly climb towards the local maximum in likelihood space. This problem, of only promising a local maximum, is a shortcoming of the EM method, but it is not a critical problem in the case of weather images because we can initialize the GMM near a "good enough" solution.

In the case of weather images, we do know that contiguous pixels "should" belong to the same Gaussian. We can take advantage of this spatial coherence to place the initial mixture components. The pixels in the image with valid data values are grouped into regions consisting of contiguous pixels. These pixels are then arranged so that all the pixels in a region are listed together. The carefully arranged list of pixels is broken into K equal parts where K is the desired number of Gaussian components. Each pixel gets a weight of one for "its" Gaussian component and zero for all other components i.e. if a pixel falls into the kth group, the weight is one for the kth component and zero for all other components.

Thus, the initial condition consists of a number of Gaussian fits so that separate regions

will tend to be fit to a Gaussian. Relatively large regions will be fit in parts to Gaussians. From this initial point, the hill climbing approach of the EM method finds the best possible fit. However, because the EM method is only a local optimization method, there may be a better solution elsewhere but it may not be reached.

d. Parameters of the GMM

The GMM is completely specified by the following parameters: π, μ_x, μ_y, σ_x, σ_y and σ_{xy} for each of the K Gaussian components of the GMM. Recall, however, the GMM was defined so as to sum to 1, and that the EM method optimized the likelihood of the parameters given the *positions* of the pixels (and not the intensity). Thus, two minor changes have to be made to the GMM procedure explained above:

i. The total intensity associated with all the pixels in the image is stored and this value,A, is used to scale the GMM so that the image intensities can be recreated i.e. the GMM equation is modified to be:

$$G(x, y) = A \sum_{k=1}^{K} \pi_k f_k(x, y) \tag{10}$$

ii. Because the EM method does not cater to the intensity, the more intensive locations are repeated several times. This is done by creating a cumulative frequency distribution (CDF) of the pixel values in the image and using a pixel's location m times where m is given by:

$$m = 1 + \gamma \; round(\frac{CDF(I_{xy})}{freq(I_{mode})}) \quad \forall \; I_{xy} < I_{mode} \tag{11}$$

where I_{mode} is the intensity corresponding to the most frequent quantization interval in

the histogram of intensities used to compute the CDF. Pixel locations with intensities lower than I_{mode} are used only once. It is apparent that if the correction factor, γ, is zero, then pixels are not repeated and as *gamma* is increased, higher intensity pixels are repeated more often. The results in this paper, unless explicitly stated otherwise, all use $\gamma = 1$.

The need for, and the effect of, this intensity correction can be illustrated using the artificial dataset shown in Figure 2. Without intensity correction (See Figure 2b), the GMM fit simply tries to get all the non-zero pixel locations correct and the resulting GMM fit is simply a symmetric ellipse. With low values of γ (See Figure 2c), because there are many more low-intensity pixels than high-intensity pixels, the GMM fit is dragged only slightly towards the higher intensity values. On the other hand, when the higher intensity pixels are heavily emphasized (See Figure 2e), there are many more high-intensity pixels in the fit and therefore, several components of the GMM are expended towards getting the high-intensity locations correct. In this paper, we use the moderate value of $\gamma = 1$ because it appeared to work best on real precipitation forecast fields.

e. Error Measures

Given two Gaussian components, one from the forecast field and one from the observed field, it is possible to compute translation, rotation and scaling errors from the parameters of the two components (how corresponding Gaussians are identified is described in Section 2f).

The translation error, e_{tr}, is the Euclidean distance between their means:

$$e_{tr} = \sqrt{(\mu_{xf} - \mu_{xo})^2 + (\mu_{yf} - \mu_{yo})^2} \tag{12}$$

where the subscripts f and o correspond to the forecast and observed fields respectively.

The rotation error, e_{rot}, can be computed from the two covariance matrices since the first eigen vector of a covariance matrix represents the direction of maximum variance (this is the key idea underlying Principal Components Analysis, for example). Once the eigen vectors of the two covariance matrices are computed, the dot product of the eigen vectors yields the cosine of the angle between them. Hence, the rotation error (in degrees) can be computed as:

$$e_{rot} = \frac{180}{\pi} cos^{-1}(v_f.v_o) \tag{13}$$

where v_f and v_o are the maximum-variance eigen vectors of the covariance matrices (Σ) of the forecast and observed fields. As pointed out by Davis et al. (2006), however, one should be careful about using rotation error on objects that are circular. In the case of a GMM, the confidence associated with e_{rot} is low if σ_x and σ_y are nearly equal.

The scaling error, e_{sc} can be computed as:

$$e_{sc} = \frac{A_f \pi_{k_f}}{A_o \pi_{k_o}} \tag{14}$$

so that if e_{sc} is less than one, it's an underforecast and if it is greater than one, it's an overforecast.

f. Finding Corresponding Gaussians

All the error measures in the previous section are defined assuming that one Gaussian component from each field (forecast and observed) is given. In fact, there will be K Gaussian components available from each field. Therefore, these error measures are computed for each pair of Gaussian components (K^2 pairs in all) and the best match for each forecast component is selected by normalizing and weighting the three individual errors to compute an overall error. We chose the scaling factors and weights arbitrarily:

$$e = 0.3 * min(\frac{e_{tr}}{100}, 1) + 0.2 * min(e_{rot}, 180 - e_{rot})/90 + 0.5 * (max(e_{sc}, 1/e_{sc}) - 1) \quad (15)$$

In practice, they would be chosen based on the resolution of the images and the needs of the users of the forecast. For example, underforecasts and overforecasts may have different costs, as could translation errors beyond a certain threshold.

The overall forecast error is defined as the mean of the individual GMM component errors. Alternately, because the Gaussians are localized, the errors could be used as indicative of the errors in different regions of the forecast field.

g. Number of Components

The initialization procedure assumed that we needed a GMM consisting of K components. How do we know the number of components needed in the GMM?

The traditional way to estimate K is to start with 1 model and slowly increase the number of models. At each K, the log-likelihood obtained from the GMM fit is used to compute an

information criterion such as the Bayes Information Criterion (BIC) (Hand et al. 2001):

$$BIC = 2l(\theta) - 6Klog(N) \tag{16}$$

The optimal value of K is the K at which the information criterion is maximum. In effect, the fitting is stopped when the number of parameters to represent the model (μ_x,μ_y, σ_x,σ_y,σ_{xy},π for each of the K Gaussian components) starts to overwhelm the advantage gained by the increased likelihood.

We found though, the maximum number of components given by this criterion is too many for the forecast verification problem. For example, for the image shown in Figure 1, the number of components required before the BIC stops increasing is on the order of hundreds. Thus, we subjectively chose the maximum number of components to be 3 for all the cases considered.

3. Results, Analysis and Conclusions

We computed the GMM on three datasets from a verification methods intercomparison project (Gilleland et al. 2009; Ahijevych et al. 2009) that was established to improve the understanding of the characteristics of various model forecast verification methods. The goal of the intercomparison project was to provide answers to questions such as how different verification methods provide information on location errors, intensity errors, structure errors and model performance at different scales. To enable reasonable comparison, the verification methods were carried out on synthetic and real fields with known errors. The methods were also applied to a common dataset used in a subjective model evaluation ex-

periment. The results of the GMM approach on the different datasets that were created by the intercomparison project are presented below.

a. Geometric

This dataset consists of a synthetic object that is subjected to geometric transformations. We carried out GMM fitting assuming 3 components so as to keep the hand-analysis of GMM parameters manageable. For consistency, we used the normal intensity correction ($\gamma = 1$) that we employ on real-world datasets.

Even though these choices are non-ideal for this synthetic object, the GMM approach does extremely well in identifying the translation, rotation and scaling errors. The GMM fit shown in Figure 3 is a poor approximation to the synthetic object. This is because the synthetic object is unrealistic in two specific ways. First, the synthetic object has abrupt transitions between intensity levels whereas Gaussian approximations are better suited to more gradual varations. Secondly, the intensity (gamma) correction is done based on a cumulative distribution function. This works well on real-world images but does poorly on this synthetic image where the distribution function consists of just two values. Indeed, as shown in Figure 2, it is possible to obtain a better approximation to the synthetic object by using many more components (to better approximate the high gradients) and a higher value of γ (to better equalize the sparse intensity histogram).

By referring to Table 1, it may be observed that translation to the right, whether by 50 points as in geom001 or by 125 points as in geom005, is easily inferred by the change in the longitude direction of the appropriate number of pixels. Translation to the north

or south can similarly be inferred from changes in μ_y. Differences in size can be inferred quantitatively as changes in σ_x or in the amplitude,$A\pi_k$, as in geom004. Both numbers ($\sqrt{2110/128}$ and 167034/49734) indicate that the region in geom003 is about three times too big. The wrong orientation in geom004 can be inferred from the changes in σ_x and σ_y. The new object is 4 times too small in the north-south direction and 4 times too large in the east-west direction. The translation by 125 pixels can be inferred by the change in μ_x. Quantitatively, the rotation is captured by the e_{rot} of 90 degrees. When the objects become circular (as in geom003 and geom005), the rotation metric is unreliable but this is to be expected because the "orientation" of a circular object is undefined. Thus the GMM is able to capture the transformations on this synthetic dataset (except for circular objects).

If we were to rank the different synthetic forecasts by the admittedly subjective weighted error metric of Equation 15, the order is: geom001, geom002, geom004, geom003 and finally geom005. This is intuitively what one would expect.

b. Perturbed

The "perturbed" set of cases from the Intercomparison Project (Ahijevych et al. 2009) consists of observed data from the 2005 NSSL/SPC Spring Experiment described in Kain et al. (2008). The observed data were subjected to various transformations as shown in Figure 4. We carried out the fit with 3 Gaussian components, as in the case of the synthetic cases, primarily to keep the hand-analysis of GMM parameter changes tractable. We used only the top 10% of pixel values in each of the images to form the GMM fit so as to avoid contanimation by the extremely large number of low intensity pixels in this real-world image.

This adaptive threshold was 6.6 mm on the original image and higher, due to movement of pixels beyond the edge of the domain, for the perturbed images.

Here too, the GMM is able to capture the translations as shown in Table 2 for cases 1-3. Within the limits of round-off error, the differences in μ_x and μ_y match up well with the known translation errors (See also the first two columns in Figure 4). In cases 4 and 5, the translations are larger. While the GMM fits and e_{tr} point to the magnitude of the translation error, the numerical estimates are inexact because many of the pixels that were in the original fit are now off the edges of the image. The dependence of the GMM fit on these boundary pixels can be derived analytically and is given by the partial derivative of the GMM equations with respect to x and y. If pixels in the eastern part of the image are not included in the GMM fit, for example, the centroid moves to the west by an amount given by the partial derivative of Equation 6 multiplied by the number of such pixels.

Case 6 involves both translation and an overestimate of precipitation amounts – each pixel's value is multiplied by 1.5. This overestimate is captured in the amplitude ($A\pi_k$) of the Gaussian and in the scaling errors (e_{sc}s). Moreover, the translation effect is mostly independent of the amplitude effect as can be noticed by comparing the μ_x and μ_y here with those of fake003. The translation error in fake006 is not identical to that of fake003 because formerly low-intensity pixels around the boundaries of a storm system were included in the GMM fit once their intensities are multiplied by 1.5.

Finally, fake007 involves both translation and a consistent underestimate of precipitation. This is reported by the GMM as a reduction in the amplitude and in the size (σ_x is smaller and σ_y larger but the net change is towards a smaller size). Note, for comparison, that fake006 showed an amplitude increase but no increase in size. Thus the GMM is able

to parsimoniously capture all the transformations on the perturbed dataset. The underforecast is captured in e_{sc} but because the e_{sc} was defined as a ratio, the reported error (0.67, for example) does not match up with the actual transformation which was a constant underforecast of 2mm.

Ranking the different perturbed forecasts by the error metric of Equation 15 yields this order: fake001 (0.02), fake002 (0.04), fake003 (0.23), fake006 (0.31), fake004 (0.33), fake007 (0.42) and finally fake005 (0.44). Ordering forecasts in this manner is subjective as the order would change depending on the weights assigned to the translation, rotation and scaling errors and to the maximum tolerable errors in each category.

c. June 1, 2005

The third set of cases we analyzed consists of observed data and model runs from the 2005 NSSL/SPC Spring Experiment described in Kain et al. (2008). The observed data from June 1, 2005 are compared with 24 hour forecasts of one hour rainfall accumulation carried out on May 31, 2005. The GMM fits of the data and the model forecasts (from the 2CAPS, 4NCAR and 4NCEP models) are shown in Figure 5. The images cover the lower 48 states of the United States. The 4NCEP model forecast was produced at the National Centers for Environmental Prediction (NCEP) using a Weather Research and Forecasting (WRF) model whose core was a Nonhydrostatic Mesoscale Model (Janjic et al. 2005) with a 4.5km grid spacing and 35 vertical levels. The 4NCAR model forecast was produced at the National Center for Atmospheric Research using the Advanced Research WRF (ARW; Skamarock et al. (2005)) core with a 4km grid spacing and 35 vertical levels. The 2CAPS was produced

at the Center for Analysis and Prediction of Storms at the University of Oklahoma (also using the ARW core) with a 2km grid spacing and 51 vertical levels. All three forecast systems used initial and lateral boundary conditions from the North American Model (Rogers et al. 2009). The observations are from the Stage II rainfall accumulation dataset produced by NCEP (Baldwin and Mitchell 1998).

The June 1 case consists of three quite different systems: an elongated band stretching north-south in the middle of the image, somewhat weaker precipitation in the Southeast and weak, isolated storms in the Northwest. As with the "fake" cases in the previous section, we carried out the fits with 3 Gaussian components for tractability and limited the fit to the top 10% of pixel values in each of the images. The 3-component GMM fit does not capture these three events. Instead, two of the components correspond to the northern and southern sections of the elongated band and the south-eastern band. The weak, isolated cells in the Northwest are ignored in the GMM fit. As pointed out by Wernli et al. (2009), it would be advantageous to carry out this analysis on smaller domains where only one type of of meteorological system predominates. It should also be noted, from Figure 1, that higher order GMM fits do capture all these systems. We chose to use only a 3rd order fit so as to keep the hand-analysis of component parameters tractable. An automated analysis employing more components is shown in Figure 6.

The GMM coefficients are shown in Table 3. The GMM coefficients of the 2CAPS forecast (which is the same as the fake000 field in Table 2) are repeated for convenience.

The easy correspondence of GMM parameters that existed in the geometric and perturbed cases does not exist in the real model forecasts. Nevertheless, interesting conclusions can be drawn from the transformations indicated by the changes in the GMM parameters. We'll

consider the Gaussian components one-by-one.

For the first Gaussian component (corresponding to the Northcentral part of the image), all three forecasts are displaced to the north and west. The 2CAPS forecast is the least displaced – its μ_x and μ_y are closest to that of the observation and e_{tr} is lowest. The 4NCAR model run underestimates the precipitation; the 2CAPS model run overestimates it while the 4NCEP gets the intensity of precipitation nearly correct ($A\pi_k$ of 23002 vs. 22136 or a e_{sc} of 1.04). Examining the elements of the Σ_{xy} matrix, the 2CAPS forecast gets the shape wrong whereas the 4NCAR and 4NCEP forecasts get the extent correct in the north-south direction (the x direction in our right-handed coordinate system centered at the top-left of the image) but over-estimate the east-west extent.

For the second Gaussian component (corresponding to the Southcentral part of the image), all three forecasts are displaced to the north, with the 2CAPS forecast again exhibiting the least displacement. The forecasts are extremely vertical (ratio of σ_y to σ_x) whereas the observation indicates that the field should be more horizontal. The wrong orientation is captured in e_{rot}, although this error might exaggerated because the 3-member GMM fit does not adequately capture the curvature in the line. In terms of intensity ($A\pi_k$ or e_{sc}), the 2CAPS is the closest whereas the 4NCAR and 4NCEP forecasts are significant overestimates.

On the third Gaussian component (covering the Southeastern part of the image), the NCAR and NCEP model forecasts get the intensity and orientation correct but are displaced to the east. The 4NCEP also exhibits a displacement to the north. In addition, the 4NCEP's forecast is overly large in the north-south direction indicating the precipitation, even if correct in the aggregate, is spread over too large an area.

Overall, the rank of the models, based on the subjective weighting used in Equation 15, is

2CAPS (0.34), 4NCAR (0.49) and 4NCEP (0.50). At the extremely coarse scale at which the forecasts have been compared, the 2CAPS forecast exhibits the least translation, orientation and scaling errors.

If we increase the number of Gaussians, it is possible to perform the comparison at finer detail. Recall that we used 3 components in this paper only so that we could do a hand-analysis of the Gaussian components. Since even a 50-component GMM fit takes just 0.05 seconds to carry out, an automated analysis of errors can be carried out by varying the number of components from one to 50. This is shown in Figure 6. The errors plotted in that graph are the translation, rotation and scaling errors scaled according to Equation 15 i.e. the rotation error plotted there is:

$$min(e_{rot}, 180 - e_{rot})/90 \tag{17}$$

so that the errors can be averaged across components and plotted on a consistent (zero to one) y-axis. Looking at the total error graph at the bottom right of the figure, the relative rankings of the models are quite constant. The 4NCEP model exhibits the greatest errors while the 2CAPS one exhibits the least. The 4NCAR model is intermediate between these two, although at some scales (notably around 15 components), it does better than the 2CAPS model. These relative rankings are driven most strongly by the translation errors. In terms of rotation and scaling errors, the three models have comparable performance. It is also clear that the error measures are quite robust to changes in the number of Gaussian components.

d. Areas for further exploration

This paper presents a GMM approach to model verification, but is not a full-fledged verification technique. There are some unresolved questions about the GMM approach that need to be addressed in order to create a verification technique from the ideas in this paper:

i. *Association or Deformation?* In this paper, we approximated the observed and the forecast field by separate GMMs and picked out the correspondence of the parameters in the two GMMs by looking for the match with the lowest overall error. An alternative approach that would side-step the entire association problem would be to start the E-M on the forecast field with the GMM that corresponds to the observed field and observe how the GMM components get deformed. It is not known which approach is better.

ii. *Initialization of EM* The EM approach only promises convergence to a local minimum, not a global minimum. We introduced a bias towards the "known" form of the solution by organizing pixels into contiguous regions before computing the first E-step. Exploration into other algorithms for initializing the EM process may prove beneficial.

iii. *Low intensity regions* Because our GMM formulation was based on likelihood, we emphasized higher intensities by repeating the pixels at which higher intensities were present. This would have the unfortunate side effect of deemphasizing low intensity and small cells if there is a large, high intensity cell somewhere else. The intensity correction factor γ might depend on the verification problem.

iv. *Error measures* Other error measures are possible beyond the three – translation, rotation and scaling – that were defined and employed in this paper. For example, an

error metric based on size could be defined as:

$$e_{size} = \frac{\sigma_{xf} * \sigma_{yf} - \sigma_{xo} * \sigma_{yo}}{\sigma_{xo} * \sigma_{yo}} \tag{18}$$

One possible solution to the problem of low intensity regions might be to break up large spatial areas into smaller areas and then fit GMMs to them. The approach might to be fit a GMM to the entire image, then to break the image into quartiles and fit a GMM to each quartile. This process could be repeated as often as needed to create a hierarchical set of GMMs, each of which could be analyzed to obtain the forecast efficiency at the appropriate level of detail and over the appropriate spatial area. The drawback to this would be that the GMM representations would not be tied to storm morphology.

e. Summary

In this paper, we introduced the novel approach of using a Gaussian Mixture Model to verify model forecasts. We showed that the GMM approach is able to identify translation, rotation and scaling errors in forecasts. We also identified areas where this approach can be improved in order to create a robust verification method.

Acknowledgements

Funding for this research was provided under NOAA-OU Cooperative Agreement NA17RJ1227. We thank the anonymous reviewers for considerably strengthening this paper: in particular, Figures 1f, 2 and 6 came about as responses to the reviewers' questions and suggestions.

The GMM fitting technique described in this paper has been implemented within the Warning Decision Support System Integrated Information (WDSSII; Lakshmanan et al. (2007)) as part of the w2smooth process. It is available for download at www.wdssii.org.

REFERENCES

Ahijevych, D., E. Gilleland, B. Brown, and E. Ebert, 2009: Application of spatial verification methods to idealized and NWP gridded precipitation forecasts. *Weather and Forecasting*, **0 (0)**, InPress.

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

Baldwin, M. and K. Mitchell, 1998: Progress on the NCEP hourly multi-sensor u.s. precipitation analysis for oeprations and GCIP research. *2nd Symp. on Integrated Observing Systems*, Phoenix, AZ, Amer. Meteor. Soc., 10–11.

Davis, C., B. Brown, and R. Bullock, 2006: Object-based verification of precipitation forecasts. part i: Methodology and application to mesoscale rain areas. *Monthly Weather Review*, **134 (7)**, 1772–1784.

Gilleland, E., D. Ahijevych, B. Brown, B. Casati, and E. Ebert, 2009: Intercomparison of spatial forecast verification methods. *Weather and Forecasting*, **0 (1)**, DOI: 10.1175/2009WAF2222 269.1.

Hand, D., H. Mannila, and P. Smyth, 2001: *Principles of Data Mining*. MIT Press, 546 pp.

Janjic, Z., T. Black, M. Pyle, H. CHuang, E. Rogers, and G. DiMego, 2005: The NCEP WRF

model core. *21st Conf. on Weather Analysis and Forecasting/17th Conf. on Numerical Weather Predicition*, Washington, DC, Amer. Meteor. Soc., CD–ROM.

Kain, J. S., et al., 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Weather and Forecasting*, **23 (5)**, 931–952.

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

Lakshmanan, V., T. Smith, G. J. Stumpf, and K. Hondl, 2007: The warning decision support system – integrated information. *Weather and Forecasting*, **22 (3)**, 596–612.

Rogers, E., et al., 2009: The NCEP north american mesoscale modeling system: Recent changes and future plans. *23rd Conf. on Weather Analysis and Forecasting/19th Conf. on Numerical Weather Predicition*, Omaha, NE, Amer. Meteor. Soc., 2A.4.

Skamarock, W., J. Klemp, J. Dudhia, D. Gill, D. Barker, W. Wang, and J. Powers, 2005: A description of the Advanced Research WRF version 2. Tech. Rep. NCAR/TN-468*STR, National Center for Atmospheric Research, 88 pp., Boulder, CO. Available from UCAR Communications, P.O. Box 3000, Boulder CO 80307.

Wernli, H., C. Hofmann, and M. Zimmer, 2009: Spatial forecast verification methods intercomparison project – application of the SAL technique. *Weather and Forecasting*, **0 (2)**, DOI: 10.1175/2009WAF2222 271.1.

List of Tables

TABLE 1. GMM fits on synthetic images from Ahijevych et al. (2009) and the associated errors. The numbers in bold are referenced in the text. Each row refers to a Gaussian component.

Data set	Description	μ_y	μ_x	σ_y^2	σ_{xy}	σ_x^2	$A\pi_k$	e_{tr}	e_{rot}	e_{sc}	e
geom000	Original	249	203	1720	4	128	49734				
		249	203	1667	4	127	49734				
		250	203	1668	9	127	49737				
geom001	50 pts. right	249	**253**	1694	0	129	49731	**50**	0	1	0.15
		250	254	1682	4	121	49741	51	0	1	0.15
		250	253	1679	4	131	49732	50	0	1	0.15
geom002	200 pts. right	249	**404**	1612	4	126	49739	**201**	0	1	0.3
		250	403	1682	4	127	49735	200	0	1	0.3
		250	403	1760	0	129	49731	200	0	1	0.3
geom003	125 pts. right,	250	**339**	1696	9	**2110**	**167034**	136	91	**3.36**	1.68
	too big	249	340	1696	13	2048	167018	137	92	3.36	1.67
		250	341	1647	4	2021	167032	138	91	3.36	1.68
geom004	125 pts. right	249	**341**	**104**	1	**2046**	49736	138	**90**	1	0.5
	wrong orientation	249	340	101	1	2027	49729	137	90	1	0.5
		250	339	105	2	2120	49740	136	90	1	0.5
geom005	125 pts. right,	249	**355**	1678	17	**8271**	**323126**	152	90	6.5	3.25
	huge	250	356	1688	34	8203	323125	153	90	6.5	3.25
		250	356	1668	16	8265	323121	153	90	6.5	3.25

TABLE 2. GMM fits on perturbed images from Ahijevych et al. (2009) and the errors associated with the forecasts. The numbers in bold are referenced in the text.

Data set	Description	μ_y	μ_x	σ_y^2	σ_{xy}	σ_x^2	$A\pi_k$	e_{tr}	e_{rot}	e_{sc}	e
fake000	Original	176	289	1305	743	1328	26437				
		309	252	1272	482	665	26437				
		379	407	1456	3919	20490	26437				
fake001	3 pts. right	**181**	**292**	1306	743	1328	26437	6	0	1	0.02
	5 pts. down	314	255	1270	490	675	26437	6	0	1	0.02
		384	410	1456	3918	20424	26437	6	0	1	0.02
fake002	6 pts. right	**186**	**295**	1307	744	1329	26437	12	0	1	0.04
	10 pts. down	319	258	1269	496	675	26437	12	0	1	0.04
		389	414	1472	3928	20348	26437	12	0	1	0.04
fake003	12 pts. right	**195**	**299**	1206	840	1133	27101	21	178	1.03	0.08
	20 pts. down	340	261	774	578	767	34201	32	16	1.29	0.28
		416	495	1051	1900	10252	17843	95	0	0.67	0.53
fake004	24 pts. right	**212**	**311**	1059	813	1111	26527	42	0	1	0.13
	40 pts. down	354	276	1239	802	837	33773	51	9	1.28	0.31
		432	483	1347	3110	13743	17566	93	2	0.66	0.54
											contd...

fake005	48 pts. right	**250**	**335**	968	801	1121	25113	87	2	0.95	0.29
	80 pts. down	387	304	1772	1052	934	33256	94	5	1.26	0.42
		452	447	1405	4659	20003	15666	83	2	0.59	0.6
fake006	12 pts. right	**192**	**298**	1096	859	**1198**	**33338**	18	1	1.26	0.19
	20 pts. down	335	263	1178	773	829	42294	28	10	1.6	0.41
	times 1.5	412	483	1264	2538	12634	22304	83	1	0.84	0.34
fake007	12pts. right	**222**	**306**	**2355**	194	**459**	**17815**	49	140	0.67	0.48
	20 pts. down	345	258	79	162	486	20620	36	138	0.78	0.34
	minus 2 mm	409	431	755	2884	20770	15932	38	3	0.6	0.45

TABLE 3. GMM fits on observed and model forecasts from Kain et al. (2008) and the errors associated with the model forecasts.

Description	μ_y	μ_x	σ_y^2	σ_{xy}	σ_x^2	$A\pi_k$	e_{tr}	e_{rot}	e_{sc}	e
Observed	193	301	3546	841	936	22136				
	350	264	684	1218	7508	22616				
	383	309	921	2032	22181	20061				
2CAPS forecast	176	289	1305	743	1328	26437	21	151	1.19	0.22
	309	252	1272	482	665	26437	43	129	1.17	0.33
	379	407	1456	3919	20490	26437	98	6	1.32	0.47
4NCAR forecast	159	260	3134	2344	7636	16464	53	129	0.74	0.44
	277	264	3369	1607	932	39139	73	126	1.73	0.7
	379	461	1729	2840	14879	21068	152	6	1.05	0.34
4NCEP forecast	168	247	3518	747	6888	23002	60	118	1.04	0.34
	278	258	3153	906	484	43675	72	117	1.93	0.82
	405	416	3920	6740	24879	20010	109	11	1	0.33

List of Figures

6 Translation, rotation and scaling errors for 24-hour model forecasts of precipitation accumulation on May 31, 2005 indicate that the 2CAPS model run exhibits the least error and that the NCAR run is close to it in terms of performance, regardless of the number of components used in the GMM fit. The forecast fields themselves are shown in Figure 5. 39

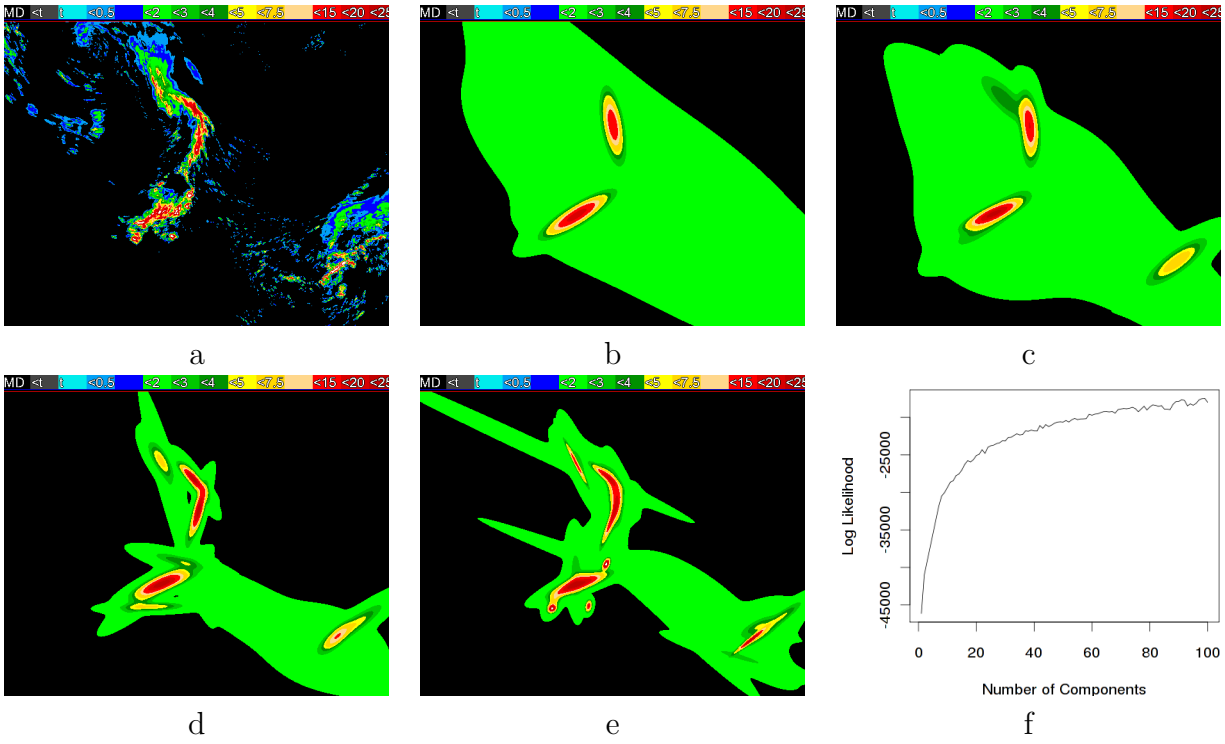


FIG. 1. Fitting a Gaussian Mixture Model to an image (a) Image being fitted: 24-hour forecast of one hour rainfall amount on May 31, 2005 from Kain et al. (2008). (b) Image recreated from a GMM with 5 component Gaussians. (c) With 10 Gaussians (d) With 20 Gaussians (e) With 50 Gaussians (f) Likelihood of the fit as the number of components is increased

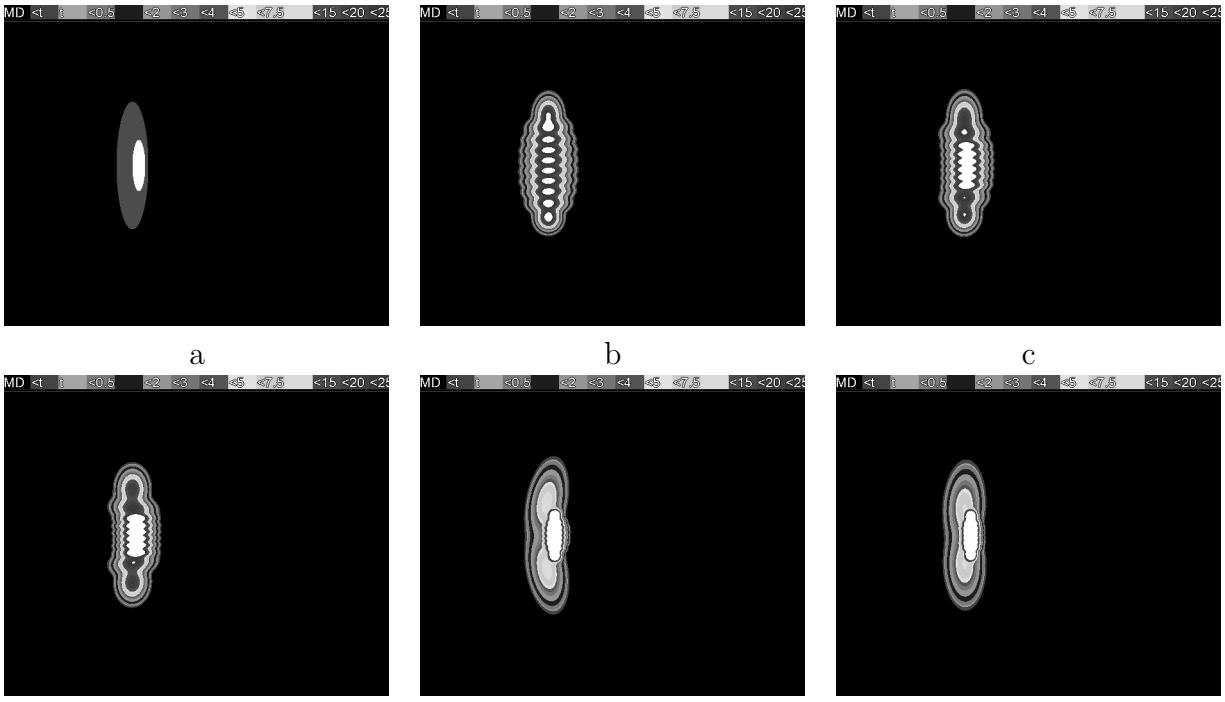
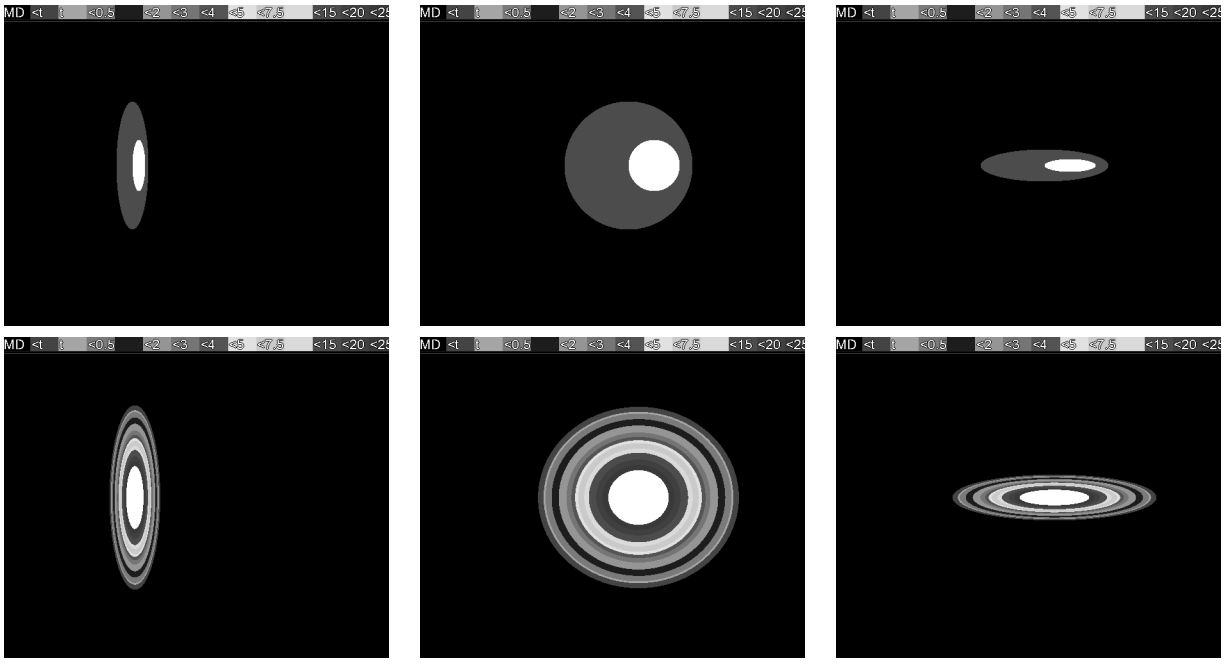
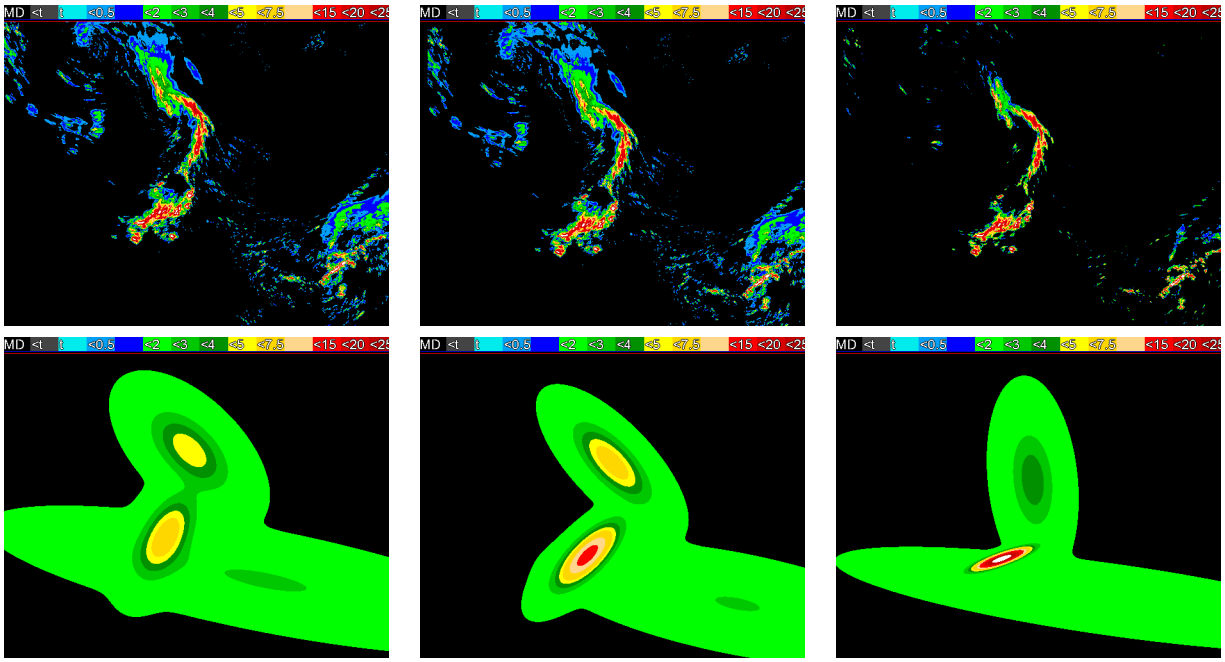


FIG. 2. Without intensity correction, the GMM will fit only the shape, ignoring the pixel values. (a) Image being fitted: Synthetic image from Gilleland et al. (2009). (b) Image recreated from a GMM with 10 component Gaussians but without any intensity correction. (c) Same as b, but with an intensity correction of $\gamma = 0.5$ (d) $\gamma = 1$ (e) $\gamma = 3$ (f) $\gamma = 5$



geom000 geom003 geom004

FIG. 3. Top row: Synthetic images from Ahijevych et al. (2009). Second row: GMM with 3 components.



fake000 fake003 fake007

FIG. 4. Top row: Perturbed images from Ahijevych et al. (2009). Second row: GMM with 3 components.

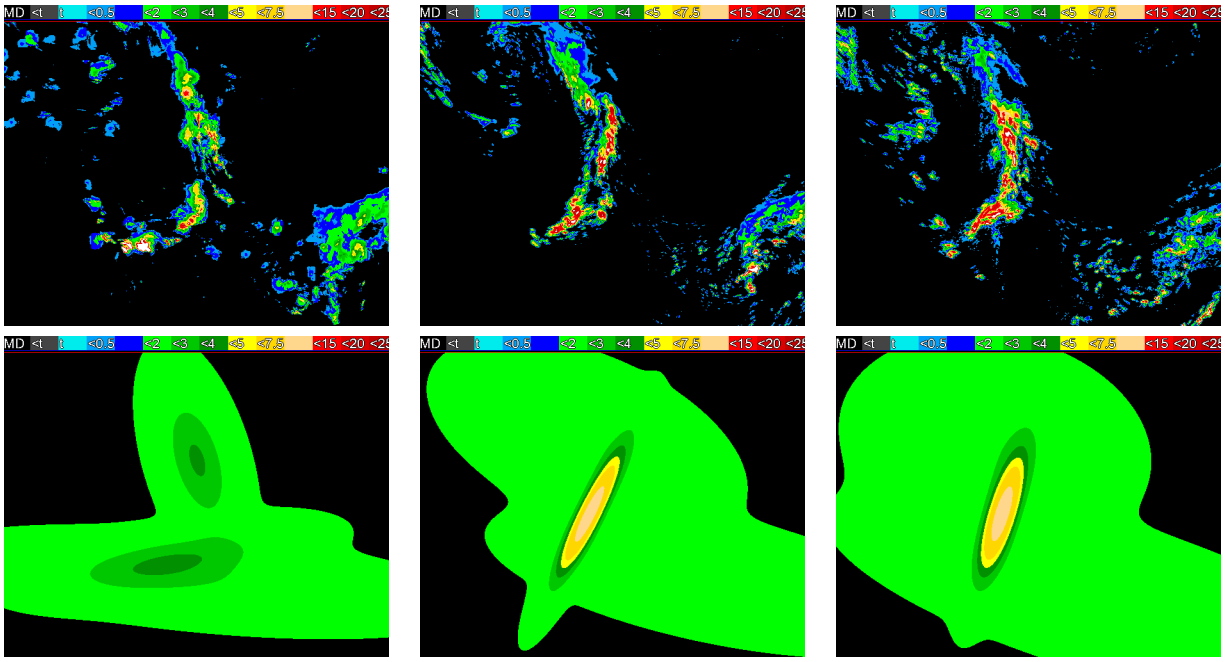


FIG. 5. Top row: Observations on June 1, 2005 and 24-hour model forecasts of one hour rainfall amount on May 31, 2005. The 2CAPS forecast field is shown in Figure 4a. Second row: GMM with 3 components.

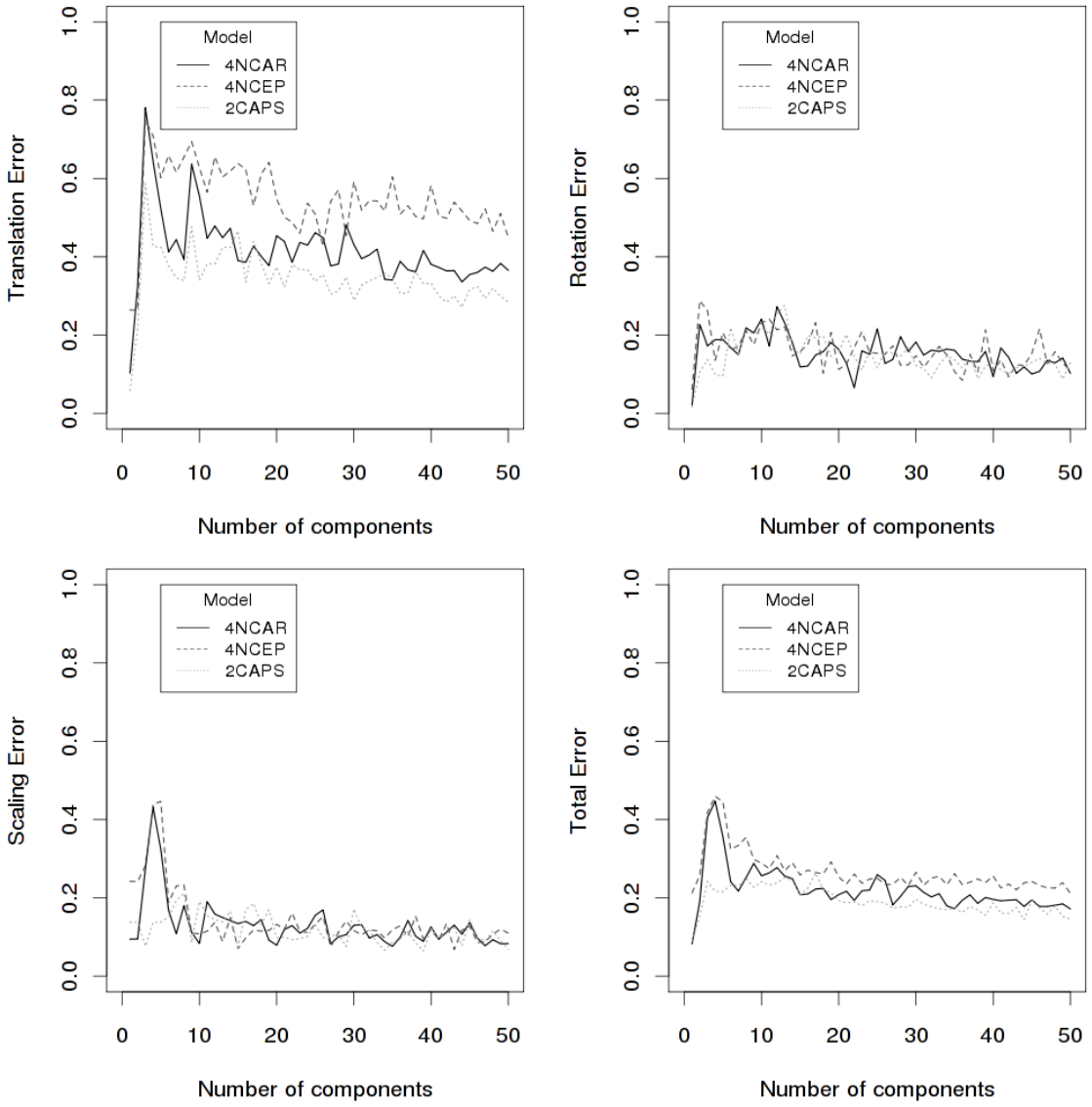


FIG. 6. Translation, rotation and scaling errors for 24-hour model forecasts of precipitation accumulation on May 31, 2005 indicate that the 2CAPS model run exhibits the least error and that the NCAR run is close to it in terms of performance, regardless of the number of components used in the GMM fit. The forecast fields themselves are shown in Figure 5.