Using Initial Condition and Model Physics Perturbations in Short-Range Ensemble Simulations of Mesoscale Convective Systems

DAVID J. STENSRUD
NOAA/National Severe Storms Laboratory, Norman, Oklahoma

JIAN-WEN BAO
Cooperative Institute for Research in Environmental Sciences, Boulder, Colorado

THOMAS T. WARNER
University of Colorado, and National Center for Atmospheric Research, Boulder, Colorado

(Manuscript received 22 January 1999, in final form 7 September 1999)

ABSTRACT

Two separate numerical model ensembles are created by using model configurations with different model physical process parameterization schemes and identical initial conditions, and by using different model initial conditions from a Monte Carlo approach and the identical model configuration. Simulations from these two ensembles are investigated for two 48-h periods during which large, long-lived mesoscale convective systems develop. These two periods are chosen because, in some respects, they span the range of convective forecast problems routinely handled by operational forecasters.

Calculations of the root-mean-square error, equitable threat score, and ranked probability score from both ensembles indicate that the model physics ensemble is more skillful than the initial-condition ensemble when the large-scale forcing for upward motion is weak. When the large-scale forcing for upward motion is strong, the initial-condition ensemble is more skillful than the model physics ensemble. This result is consistent with the expectation that model physics play a larger role in model simulations when the large-scale signal is weak and the assumptions used within the model parameterization schemes largely determine the evolution of the simulated weather events.

The variance from the two ensembles is created at significantly different rates, with the variance in the physics ensemble being produced two to six times faster during the first 12 h than the variance in the initial-condition ensemble. Therefore, within a very brief time period, the variance from the physics ensemble often greatly exceeds that produced by the initial-condition ensemble. These results suggest that varying the model physics is a potentially powerful method to use in creating an ensemble. In essence, by using different model configurations, the systematic errors of the individual ensemble members are different and, hence, this may allow one to determine a probability density function from this ensemble that is more diffuse than can be accomplished using a single model configuration.

1. Introduction

Quantitative precipitation forecasting in the summertime is one of the most difficult forecasting tasks because of the convective nature of much of the precipitation. Only limited numerical guidance is available, since numerical weather prediction models have difficulty developing and organizing convection in the correct locations and at the correct times owing to the small-scale nature of many of the features that act to initiate convection (Kain and Fritsch 1992; Stensrud and Fritsch 1994a,b; Toth et al. 1998; Buizza et al. 1999a). One approach being investigated with the hope of improving summertime quantitative precipitation forecasts (QPFs) is the use of an ensemble of model forecasts to span the space of possible outcomes.

The goal of ensemble forecasting is to predict the probability of future weather events as completely as possible (Epstein 1969; Leith 1974; Mullen and Baumhefner 1994). This is motivated by the fact that forecasts are sensitive to both small uncertainties in the initial condition (Lorenz 1963) and model error (Harrison et al. 1999), and therefore that deterministic prediction may not be reasonable. An ensemble forecast process typically begins by creating equally likely analyses of the atmospheric initial state that encompass the un-
known “true” state of the atmosphere (Leith 1974). This “cloud” of analyses represents an estimate of the probability density function (PDF) for the actual initial state, while the mean position of this cloud in phase space represents the best estimate of the actual state of the atmosphere in a least square error sense. The spread among the individual analyses is an estimate of analysis uncertainty. To create a forecast ensemble, forecasts are produced from each of the analyses, where the forecasts are assumed to represent a random sample of the PDF of the atmospheric state at various future times. If the model is perfect, the mean of the forecast ensemble is the best estimate of the true state of the atmosphere. As the forecast time increases, the forecasts continue to diverge, and eventually the mean separation between the forecasts equals the mean separation between randomly chosen atmospheric states.

Much of the research in the use of ensembles has been conducted using global models and has focused upon the synoptic-scale aspects of the forecasts at extended ranges. This is a temporal and spatial scale that is well suited for model applications, since numerical models have shown skill in the prediction of baroclinic waves. One of the main foci of ensemble studies has been the creation of the perturbed initial conditions used in creating the ensemble members. This set of model initial conditions is typically constructed by using the control analysis as the best estimate of the true initial state and adding perturbations to this analysis to define the other initial conditions. Several methods are being used to define these perturbations, including the breeding of growing modes (Toth and Kalnay 1993), singular vectors (Buizza and Palmer 1995; Molteni et al. 1996), and a Monte Carlo approach (Mullen and Baumhefner 1988; Houtekamer and Derome 1995; Du et al. 1997). A common limitation of these approaches has been the need to perform a modest number of forecasts as a result of computational constraints. This limitation has led to the desire to create a small set of initial conditions that produce the maximum dispersion (Palmer et al. 1992).

Both the breeding of growing modes and singular-vector approaches produce perturbation structures that are primarily associated with regions of baroclinic instability (Toth and Kalnay 1997), thereby producing the rapid growth of synoptic-scale features. While different sets of perturbations can be created by using different constraints (Vukicevic 1993; Toth and Kalnay 1997), the model, and its associated physical parameterizations, are tied directly to the creation of the perturbations. If the model cannot forecast a given atmospheric feature, then the perturbations created for the ensemble are unlikely to contain this feature as well.

A Monte Carlo approach adds random coherent structures to either the control initial condition (Mullen and Baumhefner 1988; Du et al. 1997) or the observations (Houtekamer and Derome 1995), where these structures represent the analysis or observational error, respectively. Houtekamer and Derome (1995) find that their approach of adding random errors to the observations produces perturbations that grow as rapidly as bred perturbations, although perturbations from singular vectors grow even faster prior to reaching their optimization period. However, ensembles created from all these approaches in general remain underdispersive when compared to observations (Buizza 1997; Toth and Kalnay 1997). These results indicate that many questions remain to be answered regarding the most useful strategy for creating ensemble initial conditions. However, another significant question that needs to be addressed is the role to be played by model physics and numerics uncertainties in the creation of ensembles.

Model physics contributions to the uncertainties are due to the imperfect representation of atmospheric processes in the model (Tribbia and Baumhefner 1988). As the time and space scales are reduced from the global to the mesoscale, the aspects of the forecast that are of most concern for verification differ. For example, at larger scales, forecast verification is primarily concerned with the general placement of synoptic-scale features, the dynamics of which are included in the equations of motion that form the foundation of the numerical model. However, at smaller scales, forecast verification is primarily concerned with the locations and amounts of precipitation and other sensible weather parameters, which are often directly affected by the assumptions used to develop the model parameterization schemes for convection and other processes. For example, Kain and Fritsch (1992) find that the simulation of a squall line is affected significantly by the trigger function that determines where and when the model parameterized convection is activated. Wang and Seaman (1997) further show that different convective parameterization schemes produce different evolutions of convective activity. When several measures of skill are examined, no one scheme is clearly better than the others. While it is well known that model parameterizations can feedback and influence baroclinic development (Gyakum 1983), the sensitivity of forecast accuracy to model parameterizations appears to be greatest for short-range forecasts, as highlighted by our inability to forecast accurately convective weather events during the summertime.

It is possible, although by no means guaranteed, that an ensemble approach can assist in providing better short-range forecast guidance (Hamill and Colucci 1997, 1998a; Du et al. 1997; Stensrud et al. 1999). Because of our present uncertainties in both the model initial conditions and model physics parameterization...
schemes, this ensemble may need to contain aspects of both initial condition and model uncertainty (Stensrud and Fritsch 1994b). Houtekamer et al. (1996) have used perturbed initial conditions with perturbed model physics, but find that, although the ensemble spread is increased, it still is too small. However, multimodel ensembles (ensembles produced using output from different models) and incorporating a stochastic representation of model error have proven to be of greater value than single-model ensembles at the medium range (Richardson et al. 1996; Buizza et al. 1999b, hereafter BMP; Harrison et al. 1999), suggesting that model uncertainty can be used advantageously in an ensemble approach.

The intent of the present study is to examine two ensemble systems, one that uses only different model physical process parameterization schemes and the other that uses only different initial conditions, for short-range simulations of mesoscale convective systems (MCSs). Unlike previous investigations using medium-range ensembles, these two methods for generating ensembles are not combined, allowing us to investigate the roles of model and initial condition uncertainty in ensemble systems separately. This should help determine whether or not model uncertainty needs to be included as part of a short-range ensemble forecasting system. Since MCSs are not well forecast with present operational models (Heideman and Fritsch 1988), yet they account for nearly 50% of the warm season precipitation in the central United States (Fritsch et al. 1986; Heideman and Fritsch 1988), two MCS events that occurred over the central United States during May and June 1985 are used to test these two ensemble methods. One event occurred in association with weak large-scale forcing for upward motion, while the other occurred in association with strong large-scale forcing for upward motion. While any conclusions drawn from only two cases must be viewed with caution, these two cases are representative of the types of events in which improved forecast guidance is needed.

Discussions of the construction of ensembles using different model physics is found in section 2, and using different initial conditions is found in section 3. An overview of the observations from the two cases, 27–29 May and 10–12 June 1985, is provided in section 4. An evaluation and verification of the simulations is found in section 5, followed in section 6 by an examination of the variance created by the two ensembles. A final discussion is found in section 7.

2. Physics ensemble

Nineteen different configurations of a mesoscale model are used to produce a 19-member ensemble of simulations starting from the identical model initial conditions. Results from Du et al. (1997) indicate that 90% of the improvement due to ensemble averaging is obtainable with ensemble sizes as small as 8 to 10 members, in agreement with Leith (1974). By using 19 members in the ensemble, we are able to investigate a wide range of model configurations and still remain within our computational limitations. The model chosen for use in this study is a hydrostatic version of the Pennsylvania State University–National Center for Atmospheric Research (PSU–NCAR) Mesoscale Model version 5 (MM5) (Dudhia 1993; Grell et al. 1994), with a two-way interactive grid-nesting procedure that allows realistic terrain features. The coarse grid domain of 96 × 96 grid points has a horizontal grid spacing of 75 km, while the inner grid domain of 73 × 73 grid points has a horizontal grid spacing of 25 km (Fig. 1). All model simulations have 23 vertical sigma levels, with the spacing of sigma levels reduced near the ground surface to better simulate the evolution of the planetary boundary layer (PBL).

One approach to generating ensembles is to use different model physical process parameterization schemes to construct various versions of a model and produce an ensemble of simulations that start from the same initial condition. A similar procedure is used by Mullen and Baumhefner (1988) and Houtekamer and Lefaivre (1997) to examine model error. We assume that all the configurations of the model are equally skillful, although this assumption has not been verified and is an important question to address in future studies. If one or more model configurations is significantly less skillful than the others, then the ensemble members should be weighted unequally to obtain the best results (Thompson 1977). However, using model output from

Fig. 1. Model outer domain at 75-km grid spacing. Also shown is the model inner domain, at 25-km grid spacing (labeled bold solid line), and the region in which calculations of root-mean-square differences (rmsd’s) are calculated in the coarse domain (unlabeled bold solid line).
seven cases and four different convective schemes, Wang and Seaman (1997) indicate that when several measures of skill are examined, no one convective scheme is better than all the others at simulating rainfall.

Three of the convective parameterization schemes evaluated by Wang and Seaman (1997) are used in the "physics ensemble" described below. While the explicit microphysical scheme (Zhang 1989) remains the same throughout all the runs, with predictive equations for cloud and rainwater below the freezing level, and cloud ice and snow above the freezing level, the convective and PBL schemes are varied, as is the moisture availability parameter (see Table 1 for the exact configurations). The following physical parameterizations are used.

**Parameterized convection:** Five different convective parameterization schemes are used within the ensemble. These include the Arakawa–Schubert (Arakawa and Schubert 1974), Betts–Miller (Betts and Miller 1986), Fritsch–Chappell (Fritsch and Chappell 1980), Grell (Grell 1993), and Kain–Fritsch (Kain and Fritsch 1990) schemes. Because of the grid spacing assumptions made in the Fritsch–Chappell and Kain–Fritsch schemes, these schemes are not used on the coarse grid. Instead, the Grell scheme is used on the coarse grid when the Fritsch–Chappell and Kain–Fritsch schemes are used on the nested grid. For the other three convective schemes, the same scheme is used on both domains simultaneously during the simulations.

**Planetary boundary layer:** Two different PBL schemes are used. The model incorporates a modified version of the Blackadar (1976, 1979) high-resolution scheme (Zhang and Anthes 1982; Zhang and Fritsch 1986) and the Burk–Thompson 1.5-order closure scheme (Burk and Thompson 1989). The Blackadar scheme is used more frequently in the ensemble owing to computational considerations.

**Moisture availability:** The standard version of MM5 uses a moisture availability parameter ($M$) to specify the partitioning between surface sensible and latent heat fluxes (Anthes et al. 1987). The $M$ values used in this study are defined by calculating the antecedent precipitation index (API) from hourly precipitation data over a 3-month period prior to the model initial time, and relating these API values to $M$ following Chang and Wetzel (1991). This control $M$ field is then varied to create ensemble members with a different partitioning of surface fluxes, since Hamill and Colucci (1998b) indicate that error growth due to soil moisture analysis errors should not be neglected. A departure from the domain-average $M$ value at each grid point is calculated and this departure is either increased or decreased by 20% to provide larger or smaller gradients in the $M$ fields, respectively. This produces a 10% mean variation in the $M$ values across the model domain, well within the range of soil moisture analysis uncertainty (Marshall et al. 1999). Although one could argue that this is more of an initial condition uncertainty, varying the values of $M$ is equivalent to perturbing the surface energy budget.

**Model initialization:** The control simulation is created by blending the National Centers for Environmental Prediction (NCEP) global analysis data with surface and rawinsonde data using the approach of Benjamin and Seaman (1985). These blended anal-

---

**Table 1. Model configurations used in the physics ensemble.** Member number 1 is the model configuration used in all the initial-condition ensemble simulations. Note that for the 10–12 Jun 1985 case, the simulations with Arakawa–Schubert failed to run, owing to numerical instability problems, resulting in an ensemble of 16 members.

<table>
<thead>
<tr>
<th>Member number</th>
<th>Convective scheme</th>
<th>PBL scheme</th>
<th>$M$ gradient value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Grell</td>
<td>Blackadar</td>
<td>Control</td>
</tr>
<tr>
<td>2</td>
<td>Grell</td>
<td>Burk-Thompson</td>
<td>Control</td>
</tr>
<tr>
<td>3</td>
<td>Grell</td>
<td>Blackadar</td>
<td>Minimum</td>
</tr>
<tr>
<td>4</td>
<td>Grell</td>
<td>Blackadar</td>
<td>Maximum</td>
</tr>
<tr>
<td>5</td>
<td>Betts-Miller</td>
<td>Blackadar</td>
<td>Control</td>
</tr>
<tr>
<td>6</td>
<td>Betts-Miller</td>
<td>Burk-Thompson</td>
<td>Control</td>
</tr>
<tr>
<td>7</td>
<td>Betts-Miller</td>
<td>Blackadar</td>
<td>Minimum</td>
</tr>
<tr>
<td>8</td>
<td>Betts-Miller</td>
<td>Blackadar</td>
<td>Maximum</td>
</tr>
<tr>
<td>9</td>
<td>Kain-Fritsch</td>
<td>Blackadar</td>
<td>Control</td>
</tr>
<tr>
<td>10</td>
<td>Kain-Fritsch</td>
<td>Burk-Thompson</td>
<td>Control</td>
</tr>
<tr>
<td>11</td>
<td>Kain-Fritsch</td>
<td>Blackadar</td>
<td>Minimum</td>
</tr>
<tr>
<td>12</td>
<td>Kain-Fritsch</td>
<td>Blackadar</td>
<td>Maximum</td>
</tr>
<tr>
<td>13</td>
<td>Arakawa-Schubert</td>
<td>Blackadar</td>
<td>Control</td>
</tr>
<tr>
<td>14</td>
<td>Arakawa-Schubert</td>
<td>Burk-Thompson</td>
<td>Control</td>
</tr>
<tr>
<td>15</td>
<td>Arakawa-Schubert</td>
<td>Blackadar</td>
<td>Minimum</td>
</tr>
<tr>
<td>16</td>
<td>Fritsch-Chappell</td>
<td>Blackadar</td>
<td>Control</td>
</tr>
<tr>
<td>17</td>
<td>Fritsch-Chappell</td>
<td>Burk-Thompson</td>
<td>Control</td>
</tr>
<tr>
<td>18</td>
<td>Fritsch-Chappell</td>
<td>Blackadar</td>
<td>Minimum</td>
</tr>
<tr>
<td>19</td>
<td>Fritsch-Chappell</td>
<td>Blackadar</td>
<td>Maximum</td>
</tr>
</tbody>
</table>
yses also are used to provide the model boundary conditions at 12-h intervals.

These different model configurations chosen to create the physics ensemble focus primarily upon different parameterizations for deep convection and the PBL, yet there is uncertainty in other aspects of the model, including cloud cover, radiation, and microphysics, that may be as large or larger. The possible roles of these other parameterizations in ensemble systems are not investigated in this study. For other model details, the reader is referred to Stensrud and Fritsch (1994b).

3. Initial condition ensemble

The second approach to producing the 19-member ensemble that is investigated in this study is to use an identical model configuration, but vary the initial conditions. The MM5 model configuration chosen for all these runs is listed in Table 1. The technique chosen for creating the various initial conditions is that used by Du et al. (1997) and originally developed by Errico and Baumhefner (1987). As in Du et al. (1997), a land–sea mask is used so that the perturbations over land are smaller than those over the ocean where analysis uncertainty is greater. These perturbed analyses are then used to create a set of 19 different initial conditions, which includes the control initial condition.

The amplitude and saturation point of these perturbation fields match the spectra obtained by Daley and Mayer (1986) from an investigation of global analyses. They demonstrate that, at scales smaller than total wavenumber 30, there is no reliable information content within the global analyses. Therefore, for horizontal scales larger than total wavenumber 30, the perturbations are characterized by white noise, and for horizontal scales smaller than total wavenumber 30 the perturbations are uncorrelated in space and have the identical spectra for the control and perturbed fields [see Mullen and Baumhefner (1989), (1994) and Du et al. (1997) for further discussion of this technique]. A vertical filter is used to project the perturbations on the first four cosine modes, in order to produce greater vertical coherence in the perturbations. The product of this procedure is a set of initial conditions that have perturbations with horizontal and vertical coherence that mimic analysis uncertainty, as found by Baumhefner (1984), Daley and Mayer (1986), and Augustine et al. (1991).

Unlike previous investigations (Mullen and Baumhefner 1989, 1994; Du et al. 1997), these unbalanced perturbations are not initialized using a nonlinear normal mode procedure. This is because, on the smaller scales, the atmosphere is not necessarily in balance. In addition, the lateral boundaries of the coarse grid are perturbed only slightly. This means that the error growth along the coarse-grid boundaries is smaller than what would typically be observed. However, the main focus of this study is the behavior of the model fields on the inner grid, which is far from the coarse-grid boundaries. Results indicate that the variances of the model fields, within the interior of the coarse grid domain3 (see Fig. 1), resulting from these initial conditions, increase with time throughout the entire 48-h period (Fig. 2). Therefore, it is believed that the neglect of significant error growth along the coarse-grid boundaries has little influence on the model behavior on the inner grid throughout the 48-h period of interest.

This procedure for generating the perturbed initial conditions is just one of many that could be used, and the results presented herein are dependent upon this choice. Other methods, such as singular vectors (Buizza and Palmer 1995; Molteni et al. 1996), bred modes (Toth and Kalnay 1993, 1997), and adding random errors to the observations (Houtekamer and Derome 1995), which produce different growth rates (Palmer et al. 1998), may yield different results.

4. Observations

Two 48-h time periods are chosen for study: 1200 UTC 27 May to 1200 UTC 29 May 1985 and 1200 UTC 10 June to 1200 UTC 12 June 1985. Both periods contain the development and evolution of large MCSs over the central plains of the United States. These two scenarios are chosen because, in some respects, they span the range of weakly to strongly forced convective forecast problems routinely handled by operational forecasters. A brief summary of each event is provided before discussing the ensemble runs.

a. 27–29 May 1985

At 1200 UTC 27 May 1985 a surface low is located in west-central Illinois, with a cold front stretching eastward from the low center (Fig. 3a). A trough connects this main low center to a deeper surface low located in far southwestern Oklahoma. A mesohigh is seen in the surface data across northeastern Oklahoma and parts of Missouri and Arkansas in association with thunderstorms that develop over this region during the early morning. These thunderstorms move slowly southeasterly over the next 12 h and progressively weaken. No other significant convection exists across the plains states until early afternoon.

At upper levels, a ridge is situated across the western plains states with a trough stretching from Iowa into Arkansas and east Texas (Fig. 3b). An objective analysis of the rawinsonde data indicates that there is subsidence across much of the northern plains states (not shown).

---

3 This region is chosen subjectively to be far enough away from the coarse-domain boundaries to limit the effects of the lack of significant error growth at the lateral boundaries, while still encompassing a large portion of the model domain.
The only region of upward vertical motion (−2.5 μb s⁻¹ at 500 hPa) is found centered over eastern Oklahoma in association with the large-scale trough and the region of cyclonic vorticity. An identical analysis conducted with data from 0000 UTC 28 May yields the same general results, indicating that there is no strong large-scale forcing for upward motion in the northern plains on this day.

As the day progresses, a dryline begins to form along the high plains stretching from southern Colorado into the Dakotas. With strong capping inversions suggested by the morning soundings (lid strength index values of 2°–5°C) and no large-scale upward motion to weaken the inversions, significant mesoscale forcing is required to initiate convection. Convection initiates along the dryline near the Wyoming–Nebraska border at 1900 UTC. The national radar composite initially indicates only a single cell that grows and intensifies over the next 2 h, producing large hail and an F3 tornado in Scottsbluff, Nebraska (NOAA 1985).

During the next 9 h, this initial thunderstorm moves southeastward and develops into a convective line that stretches southwest to northeast across northwestern Kansas by 0600 UTC. This MCS continues to move southeasterly over the next 18 h as it crosses Kansas, skirts the far northeastern corner of Oklahoma, and moves into Arkansas. By 0000 UTC 29 May, the MCS is located in far eastern Arkansas and it continues to move southeastward into Mississippi as it finally decays. This MCS produced a cloud shield large enough to qualify as a mesoscale convective complex (MCC) (Maddox 1980) and is the longest-lived MCC during 1985 over the United States (Augustine and Howard 1988).

During the 48-h time period beginning 1200 UTC 27 May, other convective activity is observed in various parts of the central United States. However, the main focus of the present study is the development and evolution of this long-lived MCS that begins in eastern Wyoming and ends in Mississippi over 30 h later.
b. 10–12 June 1985

At 1200 UTC 10 June a quasi-stationary front stretches east-to-west from Missouri into New Mexico (Fig. 4a). Several thunderstorms are seen to the north of this boundary with widespread rainshowers across much of northern Oklahoma. Surface dewpoint temperatures throughout the southern plains are above 16°C (60°F), yielding values of convective available potential energy in excess of 2500 J kg⁻¹ near the front.

An upper-level trough is beginning to exit the Rocky Mountain region and cross over the southern plains at 1200 UTC 10 June, as indicated by the region of cyclonic vorticity over Colorado (Fig. 4b). Although subsidence is indicated across much of Oklahoma and Texas at this time, an analysis of the rawinsonde data from 12 h later at 0000 UTC 11 June indicates that the shortwave trough is into Kansas and vertical motions of −2 μb s⁻¹ are present in eastern Colorado. A surface low develops in far northeastern New Mexico in association with this upper-level trough by 1800 UTC 10 June and convection develops at 2000 UTC along the frontal boundary.
boundary that extends northeastward from this surface low.

Once convection develops, it quickly organizes into an arc-shaped convective line that stretches southwest to northeast across the Texas and Oklahoma panhandles and southwestern Kansas. By 0230 UTC 11 June the convective line has grown significantly, and stretches from eastern New Mexico across the Texas panhandle, through western Oklahoma, and on into central Kansas. This MCS moves southeastward, passing the Oklahoma City area near 0400 UTC after which time the leading convective line begins to weaken. A large region of stratiform precipitation has developed to the rear of this convective line, producing rainfall across much of Oklahoma and Kansas (see Johnson and Hamilton 1988; Zhang et al. 1989).

The MCS continues to move southeastward and weaken until 0830 UTC 11 June when the leading convective line redevelops upon reaching northern Texas. The MCS now changes direction slightly and begins a more southward movement into central Texas. Two hours later, the leading convective line begins to weaken for the second time, while maintaining a general southernly movement of the entire MCS. Dissipation of the MCS occurs in central Texas by 1900 UTC.

These two MCS events represent, in some respects, the ends of the spectrum of forecast problems during the summertime. The May event occurs under a large-scale ridge axis in a region of subsidence. This MCS is initiated from small-scale forcing along the dryline, and its movement primarily reflects the internal dynamics of the system interacting with the environment. In contrast, the June event occurs in association with a shortwave trough and significant rising motion. The large-scale forcing plays a significant role in the development and movement of this MCS. Indeed, the daily operations summary of the Preliminary Regional Experiment for Stormscale Operational and Research Meteorology indicates that the May event was not forecast by the Operations Center, while the June event was forecast with a high degree of confidence (Meitin and Cunning 1986), although the details of the convective evolution were not known.

5. Ensemble verification

One approach to comparing output from the two ensemble systems is to compare the ensemble means with observations. The ensemble mean ideally should provide a better forecast than any individual ensemble member because errors in the individual forecasts cancel when averaged (Epstein 1969; Leith 1974). However, an important question to address is whether the accuracy of either of the ensemble systems is different from the other. While we can compare the differences between the ensemble systems on each given day and calculate whether or not these differences are significant, with only two case days we cannot make any statements with certainty about how these two systems would perform over many days. However, this case day comparison should provide insight into whether one system is obviously less accurate than the other and provide guidance for the role of model physics in short-range ensemble systems.

Using the ensemble mean from the physics and initial-condition simulations valid at the same time as available observations, the bias, mean absolute error (MAE), and the root-mean-square error (rmse) (Wilks 1995) of the 6-h accumulated rainfall, sea level pressure, and the temperature, specific humidity, geopotential height, and vector wind at various pressure levels are calculated from the two cases. An objective analysis procedure is used to obtain analyzed 6-h rainfall accumulations and sea level pressure at the model grid points, while the model grid points closest to the location of the each rawinsonde site are used in the calculations of the other verification data without correcting for balloon drift or duration of ascent. The values of bias, MAE, and rmse are calculated at 850, 700, 500, and 250 hPa for all rawinsonde observations available within the inner grid at the 0-, 12-, 24-, 36-, and 48-h simulation times. For each time period and variable, approximately 40 upper-air observations are included in these calculations.

The rainfall observations used are hourly data reported at over 2000 stations across the United States and collected by the National Climatic Data Center. Over the inner grid of the nested modeling system, which is the focus of this investigation, there are 906 stations that report hourly rainfall in either tenths or hundredths of an inch (2.54 mm or 0.254 mm, respectively). No measurable rainfall is assumed at active stations within the database that did not report. To smooth out the small-scale irregularities in rainfall that the model cannot reproduce, the observed hourly precipitation data are objectively analyzed to the model inner grid points. No distinction is made between the input data that use tenths or hundredths of an inch accuracy, because of the limited number of stations that report at the greater accuracy. The analysis scheme chosen uses a single-pass exponential weighting function (Barnes 1964) that captures 50% of the response for 110-km wavelengths (roughly four grid points). A comparison of this precipitation analysis with the hourly national radar composites indicates that it captures the main features of the convective activity on the two days of interest. While this gridded analysis produces greater agreement between the model and observations than does the raw station data, the general conclusions are the same when using either observational dataset.

Results of the rmse calculations show that the two

\[ \text{This is because model errors are spatially correlated, requiring the use of case day mean values in comparing the populations (see Hamill 1999). Thus, we have only two samples with which to compare the overall accuracy of the ensemble systems.} \]
ensembles appear to be equally accurate in general throughout the model simulation period for most of the variables examined (Figs. 5 and 6). As mentioned earlier, tests of statistical significance from two cases are not warranted since the errors are spatially correlated. However, as a baseline check of the two ensemble systems, these results do not show that the errors in the two systems are clearly different. Similar results are found from the calculations of bias and MAE (not shown). Both ensembles also show similar rmse growth curves, with the physics ensemble appearing more accurate for some variables, while the initial-condition ensemble is more accurate for other variables. It is not obvious from this comparison that either of the two systems are producing results that are dramatically different from the other.

a. 27–29 May 1985

One approach to displaying ensemble data is to calculate the PDF of the ensemble for a specific model parameter. Thus, the probability at which the ensemble produces rainfall of 2.54 mm or greater during a specified 6-h period is calculated. These probabilities are then compared to the gridded observations for both the physics and initial-condition ensembles.

Examination of the probability fields show that none of the initial-condition ensemble members capture the region of convective development in eastern Oklahoma and Arkansas during the first 6 h (Fig. 7a). This is the region in which convection and an associated surface mesohigh are observed at the model initial time. Since convectively produced mesohighs can be important contributors to the subsequent development of convection in numerical models (Stensrud and Fritsch 1994a,b), this suggests that none of the perturbed initial conditions replicate this type of feature. However, even without the presence of a surface mesohigh, several of the different model configurations produced convection in this region (Fig. 8a). This behavior is indicative of the different criteria used to decide when and where convection is initiated in the various convective parameterization schemes.

Both the physics and initial-condition ensembles overproduce the amount of rainfall between hours 6 and 12 across much of Oklahoma, Arkansas, and Missouri (Figs. 7b and 8b). This could be an “overshoot” effect in response to the underproduction of rainfall during the first 6 h of the simulations, and can be viewed as a spinup error in which the model convective activity initially lags the observations. However, the physics ensemble also produces convective activity along the high plains stretching from northern Colorado northward into Montana (Fig. 8b). This precipitation is indicated in the national radar summaries, although it is not clearly seen in the hourly precipitation data. Significantly, all members of the physics ensemble produce convection along the high plains, indicating a high likelihood of the development of convection, but they differ in terms of where and when the convection begins. Throughout the first 12 h, the variety of initial rainfall locations seen in the physics ensemble is significantly larger than that from the initial-condition ensemble, highlighting the importance of the different model physics to the evolution of the simulated sensible weather in the model during this time period.

Similar conclusions can be drawn from the ensemble data from 12 to 36 h (Figs. 7 and 8), where the propagation of an MCS across Kansas, northeastern Oklahoma, and Missouri is clearly seen in the observations. At all times, the physics ensemble incorporates more of the observed rainfall regions than does the initial-condition ensemble. However, the initial-condition ensemble does indicate a southeastward moving MCS as one possible solution, although it moves too slowly and is displaced northward from the observations.

While the physics ensemble incorporates more of the observed rainfall regions than does the initial-condition ensemble, this does not necessarily indicate that the physics ensemble is more skillful for this case. To evaluate model skill, the equitable threat score (ETS), bias...
score, and ranked probability score (RPS) are calculated using the 6-h rainfall totals calculated from the ensembles and the gridded observations (see the appendix). For the ETS and bias scores, the ensemble mean 6-h rainfall is calculated from all the members of each ensemble, whereas for the RPS the probability of the 6-h rainfall total exceeding a given threshold is calculated. The ETS measures the skill in predicting the area of precipitation amounts over a given threshold with respect to a random forecast (Rogers et al. 1995) and is calculated for thresholds of 0.254, 2.54, 12.7, and 25.4 mm. The bias score is the ratio of the forecast area of precipitation to the observed area of precipitation for amounts exceeding a given precipitation category (Wilks 1995). The RPS represents a scalar measure of the distance between the forecast and cumulative distribution functions (Wilks 1995) and in this study the five categories used are 1) no measurable precipitation (pp < 0.254 mm), 2) 0.254 mm ≤ pp < 2.54 mm, 3) 2.54 mm ≤ pp < 12.7 mm, 4) 12.7 mm ≤ pp < 25.4 mm, and 5) pp ≥ 25.4 mm. By using these three measures, it is possible to evaluate the ability of the two ensembles to produce an accurate precipitation simulation.

For the bias score, the initial-condition ensemble mean produces values that are closest to 1 throughout most of the simulation (Fig. 9), except during the first 6 h for the two lowest thresholds and after 42 h for totals greater than 2.54 mm when the physics ensemble produces values closest to 1. Both the individual ensemble members and the ensemble means typically have a “wet” bias, a result also seen in Du et al. (1997). The effects of ensemble averaging increase the bias score for the lowest (0.254 mm) precipitation category and reduce the bias score for the largest (12.7 mm) precipitation category. The bias scores for the two lowest precipitation categories are larger from the physics ensemble than from the initial-condition ensemble owing to the different model physics producing different evolutions of the rainfall regions, which when averaged easily produce rainfall amounts exceeding 2.54 mm over broader areas. Many of these differences in bias score are significant at the 99% confidence level that is calculated using a resampling test (see the appendix). Unfortunately, comparing the ETS values from competing forecasts is problematic if their biases are dissimilar, as shown by Mason (1989) and Hamill (1999). Therefore, prior to calculating the ETS, a constant precipitation amount is added to the initial-condition ensemble mean rainfall until the bias values from the initial condition and physics ensembles are equal (Hamill 1999).

Before examining the ETS values from the two ensembles, it is important to note that a comparison of the rainfall totals from each of the configurations over time...
Fig. 7. Probability (expressed in %) of accumulated rainfall ≥2.54 mm from the initial-condition ensemble and locations of observed rainfall ≥2.54 mm. Plots valid at the (a) 6-, (b) 12-, (c) 18-, (d) 24-, (e) 30-, and (f) 36-h simulation times. A contour interval of 20% is used for the probabilities, with the 10% line plotted as well. Observed locations denoted by the points. Model simulations are started at 1200 UTC 27 May 1985.
Fig. 8. As in Fig. 7 but from the physics ensemble.
Fig. 9. Bias scores from the initial-condition (IC) and physics (PH) ensembles vs model time for the (a), (b) 0.254-mm rainfall threshold, (c), (d) 2.54-mm rainfall threshold, and (e), (f) 12.7-mm rainfall threshold. Scores calculated for each ensemble member, plus the ensemble mean (thick solid lines) using 6-h accumulated rainfall data. Panels (a), (c) and (e) are from the initial-condition ensemble, while panels (b), (d), and (f) are from the physics ensemble. Scores for the 25.44-mm category are not shown, since for this case the model has no skill for this precipitation amount (see Fig. 11).
indicates that the total amount of precipitation generated in association with the different schemes is not the same. For each physical process parameterization scheme (convective and PBL) an average total accumulated rainfall over the inner grid is calculated for each hour of the model simulations using output from a fixed number of model configurations that include this scheme within the physics ensemble. For example, the average accumulated rainfall from the Blackadar PBL scheme is calculated from five simulations that use the Blackadar PBL scheme but have different convective schemes. The same five convective schemes are then used in calculating the average accumulated rainfall from the Burk-Thompson PBL scheme.

Results indicate that the Burk-Thompson PBL scheme produces 1.4 times as much rainfall as the Blackadar scheme, and this ratio is nearly constant through the entire 48-h period (Fig. 10a). This ratio in total rainfall is nearly as large as the value of 1.6 seen in the difference between the smallest and largest rainfall totals calculated from the five different convective schemes (Fig. 10b). This general result is the same when using output from the June physics ensemble (not shown), except that the ratios are a little smaller (~1.3). These results suggest that certain model configurations produce consistently larger rainfall totals than others, through either larger rainfall amounts or larger rainfall regions. This has a deleterious influence on the bias score of the physics ensemble and suggests that the different model configurations may have different systematic errors.

The ETS values indicate that the physics ensemble is more skillful for this weakly forced convective event than is the initial-condition ensemble when using the ETS calculated from the ensemble mean precipitation (Fig. 11), although neither ensemble is especially skillful for this case. The statistical significance level of the differences in skill scores is calculated using a resampling technique (see the appendix). We find that the physics ensemble is more skillful than the initial-condition ensemble at the 99% confidence level for the 0.254- and 2.54-mm rainfall thresholds for nearly half of the 6-h intervals (Fig. 11). The initial-condition ensemble is the significantly more skillful one for only two time intervals at the lowest rainfall threshold of 0.254 mm. At the rainfall threshold amounts greater than 12.7 mm, neither of the ensembles has any skill for this event. Also note that there are a large number of individual simulations within the physics ensemble that produce higher values of the ETS for the 0.254- and 2.54-mm thresholds, suggesting that the physics ensemble is providing information not found in the initial-condition ensemble. Each individual member of the physics ensemble also produces at least one 6-h precipitation total that is better than the ensemble mean, suggesting that none of the model configurations are substantially inferior to the others.

The additional information provided by the physics ensemble can be seen more easily by tracking the centroids of the regions of larger accumulated precipitation (>5 mm precipitation over 3 h) from the two ensembles. These tracks of modeled convective regions (Fig. 12) clearly show the greater diversity in model solutions found in the physics ensemble. The evolution of convection associated with the various members is quite different. While all the runs produce regions of organized convection that move in a general eastward direction, there are significant differences in the paths and lifetimes of the convective regions. The initial locations of convective development vary from Colorado to Wy-
FIG. 11. Equitable threat score vs model simulation time from the (a), (c), (e) IC and (b), (d), (f) PH ensembles. Scores calculated from each ensemble member, plus the ensemble mean (thick solid lines) using 6-h accumulated rainfall data. The solid square along the x axis indicates that the significance level of the differences in the ETS values for the ensemble mean is greater than 99%. Scores for the 25.4-mm category are all zero and are not shown.
Fig. 12. Tracks of the simulated MCSs from the (a) initial-condition and (b) physics ensembles during the 48-h period beginning 1200 UTC 27 May 1985. Tracks subjectively determined from 3-h model output. Observed MCS track, derived from radar data, shown in gray.

Calculations of the RPS indicate that the physics ensemble is again more skillful than the initial-condition ensemble for five out of eight 6-h periods at the 99% confidence level (Table 2). Unlike the ETS, the RPS is not affected by the differences in the bias scores of the two ensembles (see Wilks 1995). The results from the calculations of ETS and RPS suggest that, when the large-scale forcing for upward motion is weak, the assumptions contained within model physical process parameterization schemes play an important role in the evolution of the model simulations of rainfall and can be used advantageously in an ensemble system. We now turn our attention to an event where the large-scale forcing for upward motion is strong.

b. 10–12 June 1985

Since the June event is associated with much stronger large-scale forcing for upward motion, it might be expected that the behavior of the two ensembles would be different. All of the simulations produce a squall line in the Oklahoma and Texas panhandles with a southwest to northeast orientation to the line. Since there is a distinct midlevel short wave in the model initial conditions (Zhang et al. 1989), it appears that the large-scale signal is strong enough that all the ensemble members produce convection at near the same time and the same location. This robustness in the solutions may indicate a higher level of confidence that this development of convection is correct, although the relationship between skill and spread has not been seen clearly in previous ensembles over many more cases (Hamill and Colucci 1997, 1998a; Stensrud et al. 1999).

The evolution of the ensemble PDFs for rainfall ≥2.54 mm shows that the initial-condition ensemble captures most of the region of rainfall over the southern plains during the first 6 h, while also missing a large fraction of the rainfall region to the north (Fig. 13a). Similar to the May case, a larger variety of solutions is seen in the physics ensemble over the southern plains, whereas the region of rainfall observed over the northern plains is almost entirely missed (Fig. 14a). By 12 h there is a notable difference in the ensembles when compared with the results from the May case. The variations in rainfall seen in both ensembles is large by 12 h for the June event (Figs. 13b and 14b), and the initial-condition ensemble is capturing a larger fraction of the observed precipitation regions. It is during this second
Fig. 13. As in Fig. 7 but for model simulations beginning 1200 UTC 10 Jun 1985.
Fig. 14. As in Fig. 8 but for model simulations beginning 1200 UTC 10 Jun 1985.
6-h period that the major MCS begins to develop in far western Oklahoma.

The movement of the MCS can be clearly seen by following the region of observed rainfall southeastward during the next 24 h (Figs. 13 and 14). The ensemble PDFs show that this southeastward movement is captured in the model simulations, although the initial-condition ensemble is slow in moving the MCS. Indeed, the probabilities from both ensembles are generally higher than those found for the May case, suggesting that the simulations have better agreement when the large-scale forcing is strong. However, the squall lines produced by the ensemble members diverge after 18 h. Approximately half of the simulations from both ensembles maintain a southwest to northeast aligned squall line that moves southeastward across Oklahoma and into northeastern Texas. The other simulations have the squall line develop a distinct southward movement, with the squall line moving from central Oklahoma into north-central Texas. This suggests that the evolution of the MCS after it passes into central Oklahoma may be less certain, and at least two very different outcomes are possible. After 36 h a second zone of convection develops in the northern plains (not shown), and although this convection is captured in the ensembles, both ensembles also have phase errors in the placement of this convective region.

The results from the physics ensemble suggest that, while the large-scale features help to focus the initial region of convective development, the internal parameters in the physical process parameterization schemes more strongly influence the evolution of convection as the simulation time increases. Considering the significant influence that convective downdrafts have on mesoscale model simulations, as discussed by Zhang et al. (1989), this result is not surprising. Once the initial convection develops in the model simulation, the internal dynamics of the convective scheme and the resolvable-scale microphysical processes take over and help determine the subsequent evolution of the MCS.

To evaluate the skill of the ensembles for this case, the ETS, bias score, and RPS are calculated. Results again indicate that the bias score from the physics ensemble is typically larger than that from the initial-condition ensemble (Fig. 15). However, the difference in the bias scores is not as large as seen in the weakly forced event for the lower precipitation categories. However, for the 12.7-mm threshold, the bias scores from the two ensembles are very different, and the increase in the ensemble mean bias with increasing rainfall is opposite to that seen for the May case. In contrast with the May MCS event, the model simulations for this case do not appear to significantly increase the areal coverage of the rainfall region, but instead overpredict the rainfall amounts. This behavior leads to the smaller bias scores for the lower precipitation thresholds and to the larger bias scores for the higher precipitation thresholds.

To calculate the ETS, a constant precipitation amount is again added to the initial-condition ensemble mean precipitation to produce a bias score equal to that calculated from the physics ensemble (Hamill 1999). Although the ETS values for this June case (Fig. 16) are larger than those from the weakly forced May case (Fig. 11), the results are more mixed, with each ensemble being more skillful than the other at different time periods and threshold amounts. Both ensembles have skillful simulations for the 12.7-mm threshold out to 24 h, with each ensemble being more skillful than the other for two of the four 6-h periods. Neither ensemble is skillful for the 25.4-mm threshold, owing in part to the limited number of observations available for a single case.

In contrast with the May case, the values of RPS indicate that, except for one 6-h period, the initial-condition ensemble is more skillful than the physics ensemble at the 99% level throughout the 48 h (Table 3). These combined results indicate that using both different model physics parameterizations and different initial conditions are potentially of value as methods for generating perturbations in ensembles. It appears that using different model physics in an ensemble is particularly important for weakly forced large-scale environments, whereas using different model initial conditions may be a more successful approach for strongly forced large-scale environments. However, it is clear that the information produced by a model physics ensemble is important even in strongly forced events. More cases need to be examined in detail for these tentative conclusions to be verified, but these results are consistent with previous case study investigations (Stensrud and Fritsch 1994a,b; Du et al. 1997).

6. Creation of variance

The general inability of medium- and short-range ensembles to reproduce the spread seen in the observations (Buzzi 1997; Hamill and Colucci 1997, 1998a; Stensrud et al. 1999) is problematic. In essence, the spread in the ensembles is less than the difference between the forecast and the validating analyses, indicating that the ensembles are not capturing the PDF of the atmosphere. Recent results from medium-range ensemble systems suggest that the inclusion of model uncertainty increases the ensemble spread and the skill of the forecasts (Houtekamer et al. 1996; Harrison et al. 1999; BMP). Thus, we proceed to examine the creation of variance in both the physics and initial-condition ensembles. This investigation should provide some insight into which technique creates variance most rapidly, since we have shown that both the physics and initial condition ensembles are equally skillful for the two cases examined. The creation of variance is an important issue for short-range ensemble systems, since we presently have little skill in predicting summertime convection.

An important aspect in the creation of variance is the
vertical distribution. Hamill (1997) shows that ensemble dispersion is typically smaller near the surface than aloft. In contrast, Mullen and Baumhefner (1988) have found that changes in model physics tend to have a larger effect near the surface. Indeed, for short-range forecasts the simulation of surface and near-surface variables that are closely related to sensible weather are arguably the most important parameters to examine.
Therefore, the variance in model rainfall and sea level pressure are used to examine the spread in surface parameters, while the variance in temperature, specific humidity, geopotential height, and the $u$ and $v$ wind components are used to examine the spread throughout the troposphere. Above the surface, the variances are calculated for the 850-, 700-, 500-, and 250-hPa levels by interpolating the model sigma coordinate data to these levels.
pressure levels for each ensemble member. Results are summarized by calculating the mean variance across only the inner grid. To test the statistical significance of the differences between the populations of the two ensembles, the pool-permutation procedure of Preisendorfer and Barnett (1983) is used.

a. Surface variables

Results using rainfall indicate that the physics ensemble creates variance in rainfall much more rapidly than does the initial-condition ensemble (Fig. 17a), while the mean-square errors (mse’s) are nearly identical. The differences in variance are significant at the 90% level out to 15 h, after which the differences are not distinguishable from one another. In contrast, the variance in sea level pressure from the initial-condition ensemble is initially larger and remains larger than the variance from the physics ensemble until 10 h (Fig. 17b). After 10 h, the values of variance from the two ensembles are statistically identical for all but two 2-h periods. This general behavior is expected, since the initial-condition ensemble has a nonzero variance by definition in all the model fields except rainfall at the model initial time. The variance in this ensemble at first decays, owing to adjustment processes in the numerical model, and then begins to grow as seen in previous studies (Mullen and Baumhefner 1988, 1994; Du et al. 1997). Rainfall, however, is not immediately affected by the perturbations associated with the initial-condition ensemble, so that both ensembles start with zero variance in this field. Because of the different assumptions inherent in the five different convective parameterization schemes used in the physics ensemble, the timing and location of convective initiation are different in each member of this ensemble. Variance is thus produced rapidly in the physics ensemble, and it requires 15 h for the variance in the initial-condition ensemble to become as large. The decrease in variance of rainfall after 25 h is due mostly to the movement of the rainfall region out of the inner grid, although the typical diurnal cycle of convective activity in the plains states also plays a role (Wallace 1975).

The variance created by the physics ensemble grows at nearly an exponential rate during the first 6–12 h, depending upon the variable examined. This growth rate is significantly different than the more linear growth of variance from the initial-condition ensemble, although different initial condition perturbation techniques would likely produce larger growth rates (Houtekamer and Dérome 1995). However, this result is consistent with the results of Harrison et al. (1999), who find that model uncertainties produce nearly double the variance of analysis uncertainties. One interpretation of our result is that, on the mesoscale, the techniques for estimation of anal-

### Table 3. RPSs from IC and PH ensembles from simulation started 1200 UTC 10 Jun 1985. Significance level calculated using a resampling test also indicated (see the appendix). Boldface numbers indicate lowest RPS value for a given time.

<table>
<thead>
<tr>
<th>Time (h)</th>
<th>IC RPS</th>
<th>PH RPS</th>
<th>Sig. level (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0.795</td>
<td>0.974</td>
<td>99</td>
</tr>
<tr>
<td>12</td>
<td>0.769</td>
<td>0.866</td>
<td>99</td>
</tr>
<tr>
<td>18</td>
<td>0.771</td>
<td>0.790</td>
<td>99</td>
</tr>
<tr>
<td>24</td>
<td>0.890</td>
<td>0.954</td>
<td>99</td>
</tr>
<tr>
<td>30</td>
<td>0.878</td>
<td>0.840</td>
<td>99</td>
</tr>
<tr>
<td>36</td>
<td>0.796</td>
<td>0.823</td>
<td>99</td>
</tr>
<tr>
<td>42</td>
<td>0.631</td>
<td>0.703</td>
<td>99</td>
</tr>
<tr>
<td>48</td>
<td>0.475</td>
<td>0.536</td>
<td>99</td>
</tr>
</tbody>
</table>
ysis uncertainty in global models lead to an underestimation of the initial-condition perturbation amplitudes. Considering how changes in small, mesoscale regions of the atmospheric initial state can influence the subsequent simulations (Stensrud and Fritsch 1994a,b; Gallus and Bresch 1997; Spencer and Stensrud 1998), this interpretation deserves attention. However, it is not clear whether simply increasing the perturbation amplitude over the entire domain is a wise approach. It may be that the perturbation amplitudes need to be regime-dependent, or focused upon specific features, for example, convective regions, in order to provide a better estimate of analysis uncertainty on the mesoscale. Regardless of the causes of the smaller variances calculated from the initial-condition ensemble, this variance comparison suggests that modifying the model physics is an efficient approach for generating variance quickly in an ensemble prediction system, in agreement with the results of Harrison et al. (1999) and BMP for medium-range ensemble systems.

b. Upper-level variables

When examining the variance created in the temperature, specific humidity, geopotential height, and wind component fields, it is seen that the physics ensemble continues to generate variance much more rapidly than does the initial-condition ensemble while both ensembles have similar values of mse (Fig. 18). The increase in mse after 36 h is due to a phase error in the movement of the large-scale trough. Results indicate that at 850 hPa for the temperature, specific humidity, and u wind component fields, the variance produced by the physics ensemble becomes larger than the variance produced by
the initial-condition ensemble within the first 6 h of the model simulations. This vertical level is within the PBL over much of the inner grid, such that the variance can be attributed to the different PBL, surface layer, and convective parameterization schemes used. The distinct diurnal cycle of the 850-hPa temperature and specific humidity variances is not seen at vertical levels that are typically above the PBL. An examination of two-dimensional plots of the variance shows that large values of variance in the physics ensemble are concentrated in specific regions, often associated with convective activity in the model, with the other areas of the grid having smaller values of variance (Fig. 19). These large variance values dominate the total variance calculated over the inner grid and lead to the differences seen between the two ensembles (Fig. 18). In essence, the physics ensemble is very proficient at creating large differences between the ensemble members over small regions, typically where convection is produced in one or more of the ensemble members.

The evolution of the variance in geopotential height (Fig. 18a) is very different from that of the other var-

![Figure 19](image-url)
variables examined that are not vertically integrated. The variance from the initial-condition ensemble initially decreases, owing to model adjustment processes, and then begins to increase after 18 h. In contrast, the model physics ensemble produces variance in geopotential height quickly during the first 13 h, but only approaches the magnitude of the variance seen in the initial-condition ensemble. The variance from the model physics ensemble never surpasses the magnitudes seen in the initial-condition ensemble, a result not seen in all the other model fields at this level.

The difference in the evolution of the variances in geopotential height as compared to all the other model variables is attributed to the different influences on the model simulations created by the two ensembles. In the initial-condition ensemble, the initial-condition perturbations have a predefined vertical structure that often maintains the same sign of the perturbations with height. Therefore, the cumulative effect of the temperature and specific humidity fields with height is much larger than their individual contributions at a single level. In contrast, differences in model physics often lead to perturbation profiles of temperature and specific humidity that change sign with height. For instance, several of the convective schemes include downdrafts in their formulation, and when activated create strong cooling in the lowest levels of the model and warming aloft. This acts to reduce the net effect on the geopotential height calculation at a single level as compared to the other variables. This also suggests that when examining output from different ensembles, geopotential height is not the most representative variable to use in the evaluations.

As pressure decreases, the variance from the initial-condition ensemble typically becomes equivalent to, or larger than, the variance from the physics ensemble (Fig. 20). Results indicate that, while the variances from the physics ensemble grow more rapidly in the first 12 h than the variances from the initial-condition ensemble, the upper-level variances from the physics ensemble typically are less than those from the initial-condition ensemble for this case. This is particularly true of the variances in geopotential height, in which the variance from the physics ensemble is roughly half the magnitude of that from the initial-condition ensemble. Also note that the variance from the physics ensemble decreases after 24 h, when most of the convective activity moves outside the inner grid.

These conclusions drawn from the May case can also be applied to the June case without substantial modification (cf. Figs. 21 and 17), even though the large-scale forcing for the June case is much stronger than that observed for the May case. However, there is an indication that the physics ensemble is better able to produce variance in the upper levels of the atmosphere for the June case than for the May case (cf. Figs. 22 and 20). This is attributed to both the larger region of convection early in the model simulations and the higher variance from the physics ensemble for the larger region of convection.

**Fig. 20.** As in Fig. 17 except for 250-hPa (a) geopotential height (m²), (b) temperature (K²), and (c) u component wind speed (m² s⁻²) for the 27–29 May 1985 event.
equilibrium levels of the convective regions in the June case as compared to the May case. Cloud-top temperatures diagnosed from infrared satellite imagery indicate coldest temperatures between $-52^\circ$ and $-60^\circ$C during the May case (cloud tops at $\sim 200$ hPa), whereas temperatures are below $-70^\circ$C during the June case (not shown), indicating a much higher equilibrium level (cloud tops at $\sim 150$ hPa). Differences in model physics are therefore able to have a larger influence in the upper levels for the June case. Note also that there is no decrease in the variance from the initial-condition ensemble after initialization in the June event (Figs. 21 and 22), owing to the stronger baroclinic forcing that rapidly grows the perturbations in the upper levels.

It must be remembered that, in this study, the model output examined is from a relatively small inner grid. The results indicate that model physics differences play a large role in local regions when all the parameterization schemes are involved, that is, where convection occurs, and have smaller influences elsewhere in the domain. Therefore, it is possible that the global effect of model physics differences may be small, owing to the intermittent nature of these events and their relatively small size, even though the local effect is large. Since the day-to-day short-range forecast process is mainly concerned with local effects, these results strongly indicate that model physics are important to consider even if their global influence is found to be less significant.

Although the two cases simulated are very different, with one case being associated with strong large-scale forcing for upward motion and the other being associated with large-scale subsidence, the variances calculated from the ensembles for the June case are larger in
magnitude than the variances calculated from the May case. This is, at first, counterintuitive, since the June case with the strong large-scale forcing appears to be much more predictable, as indicated in the ability of the forecasters to better predict this event (Meitin and Cunnin 1986), suggesting that it should be associated with smaller variances if the ensemble output is to be useful in forecasting the forecast skill. This is consistent with results found using global models, indicating that the correlation between spread and skill need not be large (Whitaker and Loughe 1998). Therefore, for regional forecasts in particular, it may be that the ensemble spread needs to be normalized with respect to an estimate of the large-scale forcing, since with stronger large-scale forcing any errors in the initial conditions and models should grow faster (cf. Figs. 20 and 22). This normalization may allow for an improved relationship between ensemble spread and forecast skill.

7. Discussion

Ensemble techniques produce the best results when model error is small compared to initial condition error (Murphy 1988; Palmer et al. 1990). Results from the initial-condition and physics ensembles suggest that model uncertainty is large (Houtekamer et al. 1996; BMP), making the practical use of short-range ensembles less than ideal from this perspective. Mullen and Baumhefner (1989) argue that this type of result indicates that modeling error likely overwhelms any initial-condition error. However, while it is clear that the procedures for producing ensembles are greatly simplified if the model uncertainties are small, which is equivalent to making the perfect-model assumption, it is not at all clear that model uncertainties cannot be used advantageously. Thompson (1977) suggests that independent numerical predictions with different biases and error statistics can be combined to yield predictions that are, on average, better than any of the individual predictions when evaluated using a large number of forecasts (e.g., Vislocky and Fritsch 1995). Predictions from an identical model using different initial conditions tend to be highly correlated, so that its systematic errors tend to move all the predictions away from the evolution of the real atmosphere and toward the model climate. With different models, or the use of significantly different model physical process schemes within the same basic numerical framework, the systematic errors from the predictions should be less correlated than when using the identical model. The use of less correlated model predictions may lead to an ensemble that captures more of the real atmospheric PDF than an ensemble that does not include variations in model physics (Persson 1996).

Results from two case studies, using both model physics and initial-condition perturbations to create 19-member ensembles of model simulations, suggest that varying the model physics is a reasonable and potentially powerful method to use in producing an ensemble, es-

![Image](https://example.com/image.png)

**Fig. 22.** As in Fig. 17 except for 250-hPa (a) geopotential height (m$^2$), (b) temperature (K$^2$), and (c) $u$ component wind speed (m$^2$ s$^{-2}$) for the 10–12 Jun 1985 event.
especialiy since there are many uncertainties in the parameterizations used in any numerical model. This result is consistent with that found by Harrison et al. (1999) and BMP for medium-range ensemble systems. It is emphasized that the variations in model physical process schemes used here are not changes to an individual parameterization scheme, but rather the use of entirely different schemes that have been developed for use in numerical models. One can argue that all the schemes used should be able to produce acceptable simulations of the two events studied. While the conclusions are based only upon output from two ensemble cases, and need to be corroborated with a larger ensemble dataset, the cases chosen span the range of operational forecast problems and, thus, likely illustrate many of the attributes that would be found from a larger dataset.

Calculations of the rmse, ETS, and RPS indicate that the model physics ensemble is more skillful at simulating rainfall than the initial-condition ensemble when the large-scale forcing for upward motion is weak. When the large-scale forcing for upward motion is strong, the initial-condition ensemble is more skillful than the model physics ensemble. This result is consistent with our expectation that model physics play a larger role in the model simulations when the large-scale signal is weak and the assumptions used within the model parameterization schemes largely determine the evolution of the simulated weather events. The sensitivity of numerical simulations to the details of the parameterization schemes in weakly forced, large-scale environments has been documented in other studies (Stensrud and Fritsch 1994a,b; Spencer and Stensrud 1998).

The approach used to generate the initial condition perturbations is the random coherent structures technique used by Du et al. (1997) and originally developed by Errico and Baumhefner (1987). This approach is a Monte Carlo–type approach where no information about the dynamics of the given synoptic setting is used in developing the perturbations. While the results indicate that these perturbations produce growing modes, it is possible that other perturbation techniques would produce larger growth rates and more skillful simulations. However, particularly for the MCS that occurs in a weakly forced, large-scale environment, it is uncertain that any of these techniques would be much more successful, since they are designed to focus upon regions of baroclinic development (Toth and Kalnay 1997). For this case, the MCS develops under a large-scale ridge, suggesting that any growing perturbations would be found outside of the region of interest.

The variances created by the two ensembles presented in this study grow at significantly different rates, with the variance in the physics ensemble being produced two to six times faster during the first 12 h than the variance in the initial-condition ensemble. Therefore, within a very brief time period, the variance from the physics ensemble often exceeds that produced by the initial-condition ensemble. This result is dependent upon the technique whereby the initial condition perturbations are generated, and these experiments need to be reproduced when more is known about how to perturb the initial conditions for short-range ensembles. However, assuming that the members of the physics ensemble are equally skillful, such that an inferior model configuration is not substantially inflating the variance, this result strongly suggests that changes in model physics are a potentially important component in creating a short-range ensemble system. This result agrees with those of Harrison et al. (1999) and BMP for medium-range ensemble systems.

We believe that the potential importance of model physics differences to the creation of ensembles needs to be explored vigorously. While the results presented in this study are based upon the output from only two cases, and need to be corroborated with larger datasets, these two cases span, in many respects, the range of operational forecast problems. Historically, work on the problem of uncertainty in numerical forecasts has focused upon initial condition uncertainty, owing in part to the use of the perfect-model assumption in many studies. While this is a very useful assumption to make in studying the behavior of ensemble systems, models will never become perfect. They will always contain some amount of error due to numerical approximations and our imperfect understanding of atmospheric processes. While this added model uncertainty complicates the conceptual model of ensemble forecasting, the influence of model uncertainty on the simulations cannot be ignored.

Since all models contain error, both individual model forecasts and model ensembles move on trajectories that diverge from the atmospheric trajectory, even assuming the model initial condition is perfect. This divergence can happen rapidly, even over short time periods, with respect to certain model variables. But different models have different systematic errors, which leads to a divergence of the model trajectories from each other when started from identical initial conditions, as is seen in the evolution of the total rainfall fields in the different physics ensemble members. By using different models, in conjunction with different initial conditions, it may be possible to increase the accuracy and usefulness of an ensemble by creating greater divergence in the ensemble trajectories than would be created by using only different initial conditions (Fig. 23). Since many studies have indicated that the variance in the atmosphere is larger than that found in the model ensembles (Buizza 1997; Hamill and Colucci 1997, 1998a; Stensrud et al. 1999), our results suggest that this deficiency may be alleviated to some extent by including different model physical process parameterization schemes, or different models (see Richardson et al. 1996; Harrison et al. 1999; BMP), in ensemble systems. This will help create variance in the ensemble very quickly and without an apparent loss in accuracy, possibly leading to an improved relationship between ensemble spread and forecast skill.
FIG. 23. Schematic representation of the evolution of members of a forecast ensemble in phase space. The “T” represents the initial and final true states of the atmosphere, “+” represents the locations of the different model initial conditions, and “EM” represents the ensemble mean location at the final ensemble forecast time. Black lines indicate the individual model trajectories produced by the identical model for all initial condition points, while the gray lines indicate the individual model trajectories produced by different models for each initial condition point. Ellipses represent an estimate of the spread in the forecasts. The thick black line is the atmospheric trajectory.

Acknowledgments. This study was supported by the National Science Foundation under Grants ATM-9424397 and ATM-9501532. The simulations were produced on the NCAR Crays and the support of both NCAR/SCD and NCAR/MMM is deeply appreciated. Daniel Fauchier and Ann McCarthy are gratefully acknowledged for their assistance in running the MM5 model and displaying the output. Thanks are also due to Dr. Steven Mullen of The University of Arizona for his assistance in porting the initial condition perturbation scheme to the MM5 system, and for many fruitful discussions. Discussions with Harold Brooks, Chuck Doswell, Dave Baumhefner, Tom Hamill, and Mei Xu also have been a great help. The PSU–NCAR mesoscale model is maintained by the efforts of many scientists at both institutions and we are grateful for all their hard work and dedication to making this model easily accessible and user friendly.

APPENDIX

Verification Measures

a. Bias score

The bias score is simply the ratio of the number of “yes” forecasts ($F$) to the number of yes observations ($O$) given a specific criterion (Wilks 1995), such that

$$\text{BIAS} = \frac{F}{O}. \quad (A1)$$

A forecast with no bias has a value of 1. In this study, the $F$ and $O$ values are calculated for each grid point using the ensemble mean and observed precipitation amounts, and calculated according to the precipitation thresholds selected.

b. Equitable threat score

The ETS measures the skill in predicting the area of precipitation amounts over a given threshold with respect to a random forecast (Rogers et al. 1995) and is calculated in this study for thresholds of 0.254, 2.54, 12.7, and 25.4 mm. The ETS is defined as

$$\text{ETS} = \frac{\text{CFA} - \text{CHA}}{F + O - \text{CFA} - \text{CHA}}, \quad (A2)$$

where CFA is the number of correctly simulated points and CHA is the expected number of points that from a random forecast, $\text{CHA} = (F/V) \cdot O$ where $V$ is the verification area. A perfect forecast occurs when $\text{ETS} = 1$, and any forecast with $\text{ETS} \leq 0$ has no skill. Forecasts with $\text{ETS} > 0$ have skill relative to a random forecast. Hamill (1999) discusses how bias influences calculations of the ETS.

c. Ranked probability score (RPS)

The RPS represents a scalar measure of the distance between the forecast and cumulative distribution functions (Wilks 1995). In this study we use five categories of 1) no measurable precipitation ($\text{pp} < 0.254$ mm), 2) $0.254$ mm $\leq \text{pp} < 2.54$ mm, 3) $2.54$ mm $\leq \text{pp} < 12.7$ mm, 4) $12.7$ mm $\leq \text{pp} < 25.4$ mm, and 5) $\text{pp} \geq 25.4$ mm to calculate the RPS. If we define $y_i$ as the probability of precipitation forecast for the $i$th precipitation category, and $o_i$ as the observation, then

$$\text{RPS} = \sum_{m=1}^{J} \left( \frac{\sum_{i=1}^{m} o_i}{\sum_{i=1}^{m} y_i} \right)^2, \quad (A3)$$

where $J$ is the number of categories, which is 5 for this study.

The RPS is sensitive to distance and penalizes forecasts for having higher event probabilities for categories removed from the actual outcome (Wilks 1995). It is also a strictly proper score. A perfect forecast has an RPS = 0, with larger values of RPS denoting less skillful forecasts. The maximum value of RPS is $J - 1$, or 4 for this study.

d. Resampling technique

In order to test the statistical significance level of the two ensembles compared in this study, a resampling technique is used. This technique combines all the en-
ensemble members produced from both techniques (physics and initial condition) into one large dataset. Two ensembles of equal size are then created by randomly selecting the ensemble members. Skill scores are then calculated for these new ensembles and the differences in the skill scores stored. A thousand different ensembles are created in this random fashion and the resulting differences are used to compute the probability of the skill score differences to attain the value calculated from the physics and initial-condition ensembles. This is the significance level of the differences on a given case day. These results cannot be generalized to other case days.

REFERENCES


———, and ———, 1992: The role of the convective “trigger function”


