The Impact of Initial Condition Uncertainty on Numerical Simulations of Blocking

PAUL A. NUTTER* AND STEVEN L. MULLEN
Institute of Atmospheric Physics, The University of Arizona, Tucson, Arizona

DAVID P. BAUMHEFNER
National Center for Atmospheric Research, Boulder, Colorado

(Manuscript received 25 March 1997, in final form 6 November 1997)

ABSTRACT

The impact of initial condition uncertainty (ICU) on the onset and maintenance of eastern North Pacific blocking is examined within the framework of a general circulation model (GCM) and the perfect model assumption. Comparisons are made with the contrasting zonal flow regime. Twenty-member ensembles of perturbed simulations are run out to 15 days for the zonal flow, and for blocking at lead times of 8, 4, 2, and 0 days.

Blocking occurs in 95% of the 0-day lead simulations and declines monotonically to 65% for the 8-day lead simulations. The uncertainty in the exact time of onset among those simulations that form blocks also increases with lead time. The synoptic-scale features in both the blocking and zonal ensembles saturate, relative to climatological variance, and decorrelate (anomaly correlation coefficient < 0.5) by 6 days. The planetary-scale features, however, maintain skill relative to climatology beyond 10 days. The zonal simulations are generally the first to saturate and decorrelate, followed by simulations of blocking maintenance (0-day lead) and onset (2-, 4-, and 8-day lead), respectively. Thus, initial flows that project negatively (zonal flows) on the GCM's Pacific–North American teleconnectivity pattern are more sensitive to ICU, and thus are less predictable than positive (blocking flows) projections.

While the results for this study demonstrate that error growth due to ICU ultimately imposes limits on the predictability of blocking, they also suggest that skillful ensemble predictions of transitions to a blocked state are possible at long lead times if the model error is held to a minimum. The majority of the perturbed simulations make the transition into a blocked state with an associated sustenance of skill even after the loss of skill in the synoptic-scale waves. The results are consistent with the hypothesis that the planetary-scale waves may need to be preconditioned toward the formation of blocking events. They also may, in part, help explain the poor performance of operational models in forecasts of blocking onset.

1. Introduction

It has long been recognized that atmospheric blocking is capable of dominating the weather and climate of a region for time periods ranging from a week up to an entire season. Consequently, efforts are continually being made to evaluate and enhance the accuracy of blocking forecasts in medium-range weather prediction models. Experience indicates that the easiest forecasts are those that begin with a block already in the initial conditions; for example, forecasts of blocking maintenance (Simmons 1986; Tibaldi and Molteni 1990; Bengtsson 1981). A more difficult problem is to forecast accurately the onset of blocking at a lead time of several days. In particular, Anderson (1993) and Tibaldi and Molteni (1990) demonstrate that Northern Hemisphere blocking frequency is underestimated in medium-range forecast models with lead times greater than a few days. Similarly, Tracton et al. (1989) discuss individual cases in which the National Meteorological Center's (NMC, now known as the National Centers for Environmental Prediction) medium-range forecast model failed to capture the onset of blocking given a lead time of just a few days. More recently, Colucci and Baumhefner (1996) investigated the errors encountered by forecast ensembles in predicting the transition to an observed block. Although model climatologies of blocking are improving, its prediction remains a difficult task for operational models (Tibaldi et al. 1995).

When attempting to understand these difficulties with forecasts of blocking, one must always keep in mind that the state of the atmosphere shall never be known exactly. The growth of small errors in the initial state,
so-called intrinsic errors, ultimately limits deterministic predictive skill (Thompson 1957; Lorenz 1963, 1965). For this reason, it is important to assess the impact of intrinsic errors on the predictability of blocking onset and maintenance in medium-range forecast models.

Operational prediction is limited by both initial condition uncertainty (ICU) and modeling error. Because of the nonlinear nature of error growth, it is virtually impossible to separate unambiguously these two error sources using only historical archives of deterministic forecasts from the operational centers. Instead, intrinsic error growth may be isolated by generating perturbed simulations from a long-term climate model integration. The idea is to use the same model to reproduce a given control simulation starting from several different, yet plausible initial conditions. Since the values defining the initial state of the atmosphere are not known with total certainty, they may be described using a probability density function (PDF). Although the shape of the initial PDF is not known exactly, it can be estimated by generating an ensemble of different initial states that are consistent with estimates of observational and analysis errors. Simulations that are run from each of the different initial states will diverge from one another due to the inherent nonlinearity of the atmospheric instabilities being modeled. By evaluating the divergence, or spread, of the simulations, a measure of uncertainty about the control simulation is obtained. These perfect model experiments are suitable for use with any model that adequately captures the variance of the observed atmosphere at the appropriate spatial and temporal scales. It is important to note that the perturbed simulations are not referred to as forecasts because the control data are not analyses of the true atmosphere. Additional details behind the methodology of perfect model simulations are discussed by Leith (1974).

The objectives of this study are to isolate and examine the impact of ICU on the predictability of selected blocking simulations and benchmark those results against the predictability of a contrasting zonal flow regime in the absence of model error. This paper is organized as follows. Section 2 describes the experimental methods used in this study, including descriptions of the dataset, case selection, and the application of perturbations to the initial state. In section 3, the results of the ensemble simulations are evaluated in terms of blocking frequency, spatial error distribution, ensemble-averaged skill, and ensemble spread. Representative examples of perturbed simulations are then provided to help with the interpretation of the synoptic situations associated with these statistical results. Conclusions are offered in section 4.

2. Experimental design

a. Model and control data

The general circulation model (GCM) used for this study is the National Center for Atmospheric Research's (NCAR) Community Climate Model, version 1 (CCM1). The CCM1 is a sigma coordinate, global spectral GCM that utilizes 12 vertical levels and a triangular 42 truncation (T42). At this resolution, the transform grid is spaced approximately 2.8° apart.

The CCM1 includes the following parameterized physical processes: convective and stable precipitation; turbulent fluxes of sensible and latent heat; frictional heating; clear-sky and cloudy-sky radiative transfer due to both shortwave and longwave components; dependence of surface albedo on solar zenith angle; and interactions with subgrid-scale motions through diffusion. In addition, sea surface temperature and sea ice follow a seasonal trend throughout the simulation. The available surface hydrology is not utilized for this study. Details of the CCM1 are described in Williamson et al. (1987) and Hack et al. (1989).

Using the above specifications, the CCM1 is initialized with analyses from 1 October 1975, integrated for 2050 days (5.6 years), and stored in NCAR’s public archives. The control dataset for this study is obtained by extracting 0000 UTC, Northern Hemisphere 500-mb geopotential height fields from the winter period (15 November–15 March). The mean and standard deviation of this database (Figs. 1a,b) compare reasonably well with those of the observed atmosphere, a fact that validates the experimental design presented here.

b. Objective definitions

Objective definitions are required to identify consistently the onset time and duration of blocking and zonal flow events within the control and perturbed ensemble simulations. While examining several previously developed anomaly based methods to identify blocking and zonal flow events, Liu (1994) demonstrates that both the amplitude and latitudinal position of height anomalies influence the character of the identified events. However, Sausen et al. (1995) point out that definitions that are constrained by amplitude and latitudinal position are somewhat arbitrary and cannot be generically applied to different modeled or observed datasets, particularly those with differing climates. Our goal is not to develop universally applicable definitions, but rather ones that clearly and consistently identify blocked and zonal flow regimes within the framework of the CCM1 model configuration used for this study. Because the same model is used to generate the control simulations and the subsequent perturbed simulations, the use of somewhat arbitrarily defined definitions is justified for this study. Anomaly threshold levels used to define blocked or zonal flows in this study (described below) may differ from those used by other authors (e.g., Horel 1985) in response to differing data and objectives. The definitions used herein consistently identify the onset time and duration of two distinctly different
Fig. 1. Climatology of the control 500-mb height fields (m) depicted in terms of (a) the mean of all winter days, (b) the standard deviation of all winter days, (c) the mean of all days composing the six selected blocking cases, and (d) the mean of all days composing the six selected zonal flow cases.

Flow regimes that synopticians would commonly associate with blocked and zonal patterns, respectively.

The objective definition developed for this work to identify blocking is based on the persistent, quasi-stationary amplification of ridges in the planetary-scale waves. A local zonal average of 500-mb heights is formed over a distance of \( \pm 45^\circ \) (total of 90\(^\circ\)) longitude relative to a given grid point. If the height at that point is more than 200 m above this zonal average height, then a blocked point is said to exist. Inspection of individual maps reveals that the 200-m criterion allows for the consistent identification of high-amplitude ridges in this CCM1 data, which often develop into persistent blocking events. Constraining the distance covered by the local zonal average to 90\(^\circ\) longitude strengthens the emphasis on high-amplitude structures in the planetary-wave field. When this definition is applied to grid points between 50\(^\circ\) and 60\(^\circ\)N, northerly extensions of the subtropical anticyclone (see Liu 1994) are avoided. In addition, negative local height anomalies through the center of the block at lower latitudes do not discount the event as it attains...
the classic "high over low" form. Following one of Rex's (1950) requirements, a block (or blocking event) is identified if a succession of blocked points persists at a grid point for at least 10 consecutive days. Similar methods for the objective identification of blocks were used by Hartmann and Ghan (1980) and Mullen (1986, 1987).

This objective definition identifies the strongest and most active blocks in the control data near 125°W. Simulated blocking cases from this region in the eastern North Pacific are attractive for use in this research due to their large magnitude and long duration. In the observed atmosphere, many studies have demonstrated that the preferred area for wintertime Pacific blocking formation lies farther west near the international date line (e.g., Dole and Gordon 1983; LejenaÈs and Ækland 1983; Blackmon et al. 1986; Tibaldi and Molteni 1990; Anderson 1993; Sausen et al. 1995; Colucci and Alberta 1996). Given the slight difference between modeled and observed blocking climatologies, it is helpful to remind readers that precise reproductions of blocking climatologies are not required in the context of these perfect model simulations and that the use of different objective blocking criteria yields differences in blocking climatologies when applied to the same dataset. In particular, it is far more important that individual cases having similar characteristics are selected from the long climate simulation in order to isolate and examine the effects of intrinsic error growth on their development in the absence of model error.

Given that substantial, long-lived blocking events are commonly found in the control data near the eastern North Pacific, a method is developed to objectively identify periods of persistent zonal flow over the same area. To ensure that zonal flow patterns are not modified by the passage of any substantial transient ridges, the region is first searched for the existence of blocked points on any given day. The identification of blocked points proceeds exactly as described earlier, except that a local anomaly threshold of 225 m is used. The previous threshold level of 200 m is designed to identify the minimum height anomaly at which an amplifying ridge is first recognizable as a developing block. By slightly increasing the threshold level, higher amplitude ridges are allowed to propagate through the zonal flow as long as they do not enhance the amplitude of the planetary-scale wave pattern. In fact, Liu (1994) points out that extended episodes of large-scale zonal flow occur in spite of large day-to-day changes imposed by transient eddies. Again, the local anomaly threshold level of 225 m is established by trial and error after successive inspection of individual maps.

If substantial transient ridges are not found in the 500-mb height field, then zonal flow patterns are identified as zonally extended areas in which heights are no more than 50 m above climatological levels. That is, zonal flow is identified if $Z - Z' < 50$ m at all grid points along the same latitude in a band at least 50° longitude wide. At any given grid point, $Z$ is the 500-mb height and $Z'$ is the seasonal cycle defined by fitting a least squares parabola to the CCM1 control data. Whereas blocked patterns are identified as local height anomalies on any given day, zonal flow is identified using height anomalies relative to climatology. Climatological anomalies are useful for the identification of zonal flow over the eastern North Pacific since a weak ridge exists in the time-mean height field over this area (Fig. 1a). Again, the method for identification is not critical for this study as long as two distinctly different flow patterns for the same geographic area are selected.

By design, individual blocking or zonal flow cases selected as control events for this study must possess similar characteristics of magnitude, duration, location, and season. In the statistical sense, they must also be independent in time. The characteristic time ($T_o$) between independent samples of a time series can be estimated (Madden 1979):

$$T_o = 1 + 2 \sum_{n=1}^{P-1} \left(1 - \frac{n}{T}\right) \gamma^n,$$

where $T$ is the number of days in the sample and $\gamma$ is the lag 1-day autocorrelation. When $T_o$ is calculated for the CCM1 control data (not shown) a characteristic time of about 10 days between independent samples appears in the eastern North Pacific region. In view of this 10-day characteristic time period and the discussion of the previous paragraphs, blocking or zonal flow events are identified that persist for at least 10 days, and begin at least 10 days after the end of the previous event.

The onset time, duration, and mean longitudinal position of blocking and zonal flow cases that meet all the foregoing objective requirements and are selected for use in this study are listed in Table 1. Each of the six selected blocking events are found near 125°W. Nine episodes of zonal flow were originally identified in the control data that extended across the eastern Pacific. However, in order to match in number the six selected blocking cases, only the longest lasting zonal flow event from each season is chosen. The mean position of the zonal flow cases is somewhat ambiguous due to its method of identification over broad areas covering at least 50° longitude. In actuality, the east and west bounds of each selected zonal flow event (not shown) cover the range of longitudes corresponding to the selected blocking events. A composite of the blocking and zonal days associated with the six cases (Figs. 1c and 1d, respectively) demonstrates that the objective definitions above successfully identify two distinct flow regimes occurring in approximately the same geographic location.

c. Perturbations and generation of ensembles

For each of the six blocking and zonal events selected in this study (Table 1), 20 different initial fields are gen-
Table 1. Starting day, duration, and mean longitude of wintertime blocking and zonal flow events identified from the 2050-day CCM1 control simulation.

<table>
<thead>
<tr>
<th>Blocking events</th>
<th>Zonal events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case</td>
<td>Onset day</td>
</tr>
<tr>
<td>1</td>
<td>0071</td>
</tr>
<tr>
<td>2</td>
<td>0096</td>
</tr>
<tr>
<td>3</td>
<td>0427</td>
</tr>
<tr>
<td>4</td>
<td>1177</td>
</tr>
<tr>
<td>5</td>
<td>1573</td>
</tr>
<tr>
<td>6</td>
<td>1893</td>
</tr>
</tbody>
</table>

Figure 2. Schematic illustration of the generation of the five simulation ensembles. Perturbations to the initial fields are represented by the shaded circles. The horizontal lines represent the 15-day forecast period and indicate the temporal displacements of the simulations relative to the onset of the control events (heavy vertical line). The figure is not intended to represent ensemble spread.

**d. Isolation of planetary- and synoptic-scale waves**

Spectral decomposition is performed to isolate the planetary and subplanetary-scale waves. Bottger (1988) demonstrates that planetary-scale flow patterns during winter are adequately represented by spectral wavenumbers 10 and lower (≤T10). Thus, in the current study we let spectral wavenumbers less than or equal to T10 represent the planetary-scale waves and the remaining wavenumbers (>T10) represent the synoptic-scale waves. Most diagnoses of differences between simulations in this study are performed on the planetary-scale waves because the synoptic-scale error variance exceeds, or saturates relative to, the climatological synoptic-scale variance by 6 days. Further discussion of synoptic-scale decorrelation is postponed until section 3c.
3. Results

a. Reproduction of events

The frequency of blocking within each ensemble is determined by counting the members in each ensemble that develop a block. The method for the objective identification of simulated blocks is the same as that described in section 2b, except that the persistence requirement is relaxed to 5 days. The original requirement of 10 days was used to identify strong, long-lived blocking cases and follows one of the requirements established by Rex (1950). However, Dole and Gordon (1983) demonstrate that persistent anomalies that last at least 5 days have a nearly constant probability of continuing for another day. The 5-day persistence requirement applied to the simulations therefore represents a less stringent criterion used to identify blocking events while maintaining a minimum persistence time. The 10-day characteristic time period originally used to select independent blocking cases from the climatological control simulation does not apply to the 15-day perturbed simulations.

The search area for blocking events in the simulations covers the same large region (135°E–45°W) to account for any major phase shifts in the simulated blocks. However, it turns out that approximately 95% of the blocks that formed in the perturbed simulations lie within ±15° longitude of the control events, and all lie within ±30° longitude. It is interesting to note that for a given zonal wavenumber \( k \), a phase shift of \( \phi_o = \pm \pi/3 \) or \( 1/6 \) of the wavelength, corresponds to an anomaly correlation of 0.5; that is

\[
\left( \frac{1}{\pi} \right) \int_0^{2\pi} \sin(kx + \phi_o) \sin(kx) \, dx = \frac{1}{2}.
\]

Given that the distance between the upstream and downstream trough axes associated with the blocking pattern of Fig. 1c has a wavelength of approximately 90° longitude, the corresponding phase shift for an anomaly correlation coefficient of 0.5 would be ±15° longitude. Since approximately 95% of the blocking events in the perturbed simulations lie within ±15° longitude of the control, excellent phase accuracy is expected between the perturbed and control simulations. Additional results for phase accuracy of the simulations are discussed in section 3c.

Figure 4 gives the results of the search for blocking events in the simulations as ensemble blocking frequencies. The figure indicates that not all the members are able to reproduce a block. For the 0- and 2-day lead ensembles, blocks lasting at least 5 days are produced in about 95% and 85% of the simulations, respectively. Blocking frequencies fall to about 70% and 65% within the 4- and 8-day lead ensembles, respectively. Hence, as the lead time prior to the onset of the control block increases, there is a tendency for the blocking frequency to decrease. These statistics indicate that ICU imposes an upper limit on the number of simulations that are able to make the transition to the blocked regime. The extent of this upper limit may fluctuate when using larger numbers of cases, other models or model formulations, different persistent flow patterns, or blocks in different regions.

Figure 4 also indicates uncertainty in the timing of

![Figure 3](image-url)
blocking onset among the perturbed simulations. At the time when the control block is first established (day 0), the frequency of blocking is generally 30% lower than what is attained 3–4 days later. Although 5% of the 2- and 4-day lead simulations and 10% of the 8-day lead simulations generate a block one day too early, most of the transitions occur after the control. Given the random nature in which the initial perturbations are generated, there is no a priori reason to suspect this intriguing behavior.

A complementary problem observed in the 8-day lead ensemble is the development of a false alarm rate prior to the onset of the control blocks. Nearly 15% of the members develop a block that persists at least 5 days that is not observed in the control run. Evidently, a nonnegligible false alarm rate can be associated with intrinsic error growth at lead times of less than 8 days.

For comparison, the occurrence of zonal flow events within the zonal flow ensemble are counted using the method defined in section 2b at durations of 5 and 10 days. In addition, the zonal flow simulations are subjected to the 5-day blocking criteria to determine if any of the perturbed zonal flow simulations managed to develop blocking events. Table 2 indicates that 63% of the simulations develop a zonal pattern that persists at least 10 days. More impressively, all but one of the perturbed members maintain a zonal pattern that persists at least 5 days. Inspection of individual maps reveals that those simulations failing to persist for 10 days are interrupted by a passing transient ridge. In seven of those simulations the zonal flow is interrupted by a ridge that develops into a 5-day block, some starting as early as day 6. This result implies a 6% false alarm rate for blocking during the zonal flow for the control simulation. The one simulation that fails to develop a 5-day zonal pattern is not among the seven that developed a block; it is continually interrupted by high-amplitude, transient ridges. As with the blocking simulations, ICU not only imposes an upper limit on the predictability of zonal flow, it also can lead to a nonzero false alarm rate for blocking by day 6.

It is of interest to compare our perfect model results to the performance of operational forecast models. Anderson (1993) and Tibaldi et al. (1995) find that the operational models tend to predict the onset of blocking later and less frequently than observed. Discrepancies between a model’s climatology and observed climatology associated with regime transitions could lead to an underprediction of blocking, even without intrinsic error growth. But our results, albeit based on a small sample, suggest that a similar error signature could exist in a perfect model at ranges shorter than 15 days. This suggests that the shortcomings of the operational models could, to some degree, also be related to intrinsic error growth. The presence of extrinsic error would only exacerbate the situation. It is not known what the climatologies of regime transitions or the sensitivities to only ICU are in the operational models, but we believe that the problem is worthy of investigation, and in fact it would be quite tractable through use of the ensemble forecasts currently being generated at several operational centers (e.g., Houtekamer and Lefairev 1996; Molteni et al. 1996; Pauley et al. 1996; Richardson et al. 1996; Tracton and Kalnay 1993).

Although ICU imposes an upper limit on the ability of models to make a transition to blocking, the blocking frequencies shown in Fig. 4 are encouraging. The climatological frequency of 5-day blocking events over the wintertime North Pacific in the CCM1 control simulation (horizontal dashed–dotted line in Fig. 4) is approximately 30%. This value is determined by counting the percentage of days in the CCM1 control data that comprise blocking events lasting at least 5 days. The ensemble blocking frequencies shown in Fig. 4 are not directly comparable to the climatological frequency since the control blocking cases are not selected at random. Nevertheless, we find the results encouraging because a majority of the perturbed simulations are able to make the transition to blocking, even with an 8-day lead time. This suggests that predictions of blocking transitions at long lead times are inherently more skillful than random forecasts selected from climatology. Results presented in later sections will allude back to this idea in greater detail.

![Figure 4](image-url)

**Fig. 4.** Frequency of blocking within the 8-, 4-, 2-, and 0-day lead ensembles (solid lines). Frequencies are displayed relative to the onset of the control block (vertical dotted line). The horizontal dashed dotted line indicates the model’s climatological blocking frequency at 125°W.

<table>
<thead>
<tr>
<th>Case</th>
<th>Zonal ≥ 10 days</th>
<th>Zonal ≥ 5 days</th>
<th>Blocks ≥ 5 days</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>19</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>13</td>
<td>20</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>16</td>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>76</td>
<td>119</td>
<td>7</td>
</tr>
</tbody>
</table>

**Table 2.** Number of simulations within each case of the zonal ensemble that produce zonal flow lasting at least 10 and 5 days, and blocks lasting at least 5 days. An ensemble of 20 members is generated for each of the six zonal flow cases.
b. Spatial distribution of predictability error growth

It is insightful to examine the character and spatial distribution of the ensemble mean errors. At every grid point, the ensemble mean-square error (mse) between the perturbed simulations, $p$, and the control simulations, $c$, can be written in terms of the total bias and the variance of the errors. Following the decomposition by Murphy (1988), the mse over $i = \{1, \ldots, 6\}$ cases (e.g., Table 1) and $j = \{1, \ldots, 20\}$ perturbed simulations associated with each case is written

$$
\frac{1}{120} \sum_{i=1}^{6} \sum_{j=1}^{20} (p_{i,j} - c_j)^2 = (\bar{p} - \bar{c})^2 + \frac{1}{120} \sum_{i=1}^{6} \sum_{j=1}^{20} [(p_{i,j} - \bar{p}) - (c_j - \bar{c})]^2.
$$

(2)

The overbar notation represents an ensemble average, where $\bar{p} = (1/120) \sum_{j=1}^{20} p_{i,j}$ and $\bar{c} = (1/6) \sum_{i=1}^{6} c_i$.

At every grid point, the square root of each term on the right-hand side of (2) represents the magnitude of the bias and error standard deviations for the ensemble simulations. These values are calculated for the planetary waves and displayed in Figs. 5 and 6, respectively. It is useful to compare these quantities in order to help understand the relative magnitude and spatial distribution of the systematic and nonsystematic errors produced by the simulations. Results from day 6 of the simulations are presented since many stages in the development of the events are represented at this time.

Figures 5 and 6 indicate that error magnitudes depend on flow type, stage of transition to blocking, and
position relative to the control ridges and troughs. For example, the largest biases (Fig. 5) occur near the centers of the troughs and ridges associated with fully developed control blocks (i.e., 0-, 2-, 4-day leads). Since the control events are selectively chosen for their extremely large height anomalies, this signature suggests that the perturbed simulations have, on average, smaller amplitudes than the control events. This interpretation is also consistent with the fact that not all the perturbed simulations developed blocking events. Comparison of Figs. 5 and 6 reveals that the bias, or systematic error, contributes much less to the mse than the random error component. The relatively small bias is consistent with the lack of model error under the perfect model assumption. It seems less likely that a similar small bias would result from ensemble forecasts of observed blocks that include model error. The largest standard deviations in the errors (Fig. 6) are on the order of 100 m and occur slightly upstream of the amplifying ridge. Note that small, random shifts in the phase and amplitude of the planetary waves may lead to tremendous localized errors. This is especially true in the vicinity of the jet stream axis, where the size of the height gradient is large for high-amplitude features such as blocks. The largest error standard deviations occur locally at the trough and ridge axes of the 2- and 4-day lead ensembles. Since day 6 of these ensembles is just a few days after

Fig. 6. As in Fig. 5 but for the standard deviation of the ensemble error. The standard deviation field is contoured with dashed lines every 30 m; shading begins at 60 m.
the onset of the control block, this variability is consistent with our earlier finding of uncertainty in the timing of blocking onset.

c. Normalized dispersion

The total mse for the perturbed simulations can be normalized by the mse obtained for the model’s climatology. A skill score (SS) can be defined as

$$SS = 1 - \frac{MSE_p}{MSE_r},$$

where $MSE_p$ is the spatial mse for the height field of the perturbed simulations on any given day and $MSE_r$ is the model’s climatological, or reference mse for that day (Murphy 1988). A perfect simulation has a score of one. When SS reaches zero, the simulation no longer has added value, or saturates relative to climatology. As Murphy (1988) demonstrates, the score can be decomposed into nondimensional components that individually describe differences in terms of phase and amplitude. The decomposition of SS may be represented schematically as $SS = (A - B + C)/(1 + C)$, where term $A$ represents the contributions to skill in terms of anomaly correlation or phase, term $B$ represents the degradation of skill due to the total bias or amplitude error, and term $C$ represents a small (negligible for perturbed simulations) correction for the variability of the control data. The accuracy of the ensemble simulations is evaluated in terms of ensemble-averaged phase error ($A$), amplitude error ($B$), and total skill (SS). Details of the SS calculations are presented in appendix A, and testing procedures for the statistical significance of differences among ensemble mean scores are discussed in appendix B. What immediately follows is an interpretation of these results based on a decomposition of SS. Unless explicitly noted, discussion is restricted to differences significant at the 1% confidence level.

1) SYNOPSIS

Figure 7 shows the SS for the synoptic scales (spectral wavenumbers $> T10$). On average, the error growth of the synoptic-scale features saturates by 6 days for all cases. Significance tests (appendix B) reveal no statistically significant differences among the ensemble mean scores during the first 6 days. Evidently the intrinsic error growth of the synoptic-scale waves is essentially equal, regardless of the planetary flow regime in which they are embedded or time prior to block formation. On this scale, rapid decorrelation of the patterns is expected due to the nonlinear growth of intrinsic errors resulting from baroclinic instabilities. Tracton (1990) and others have suggested that scale interaction processes are important in the development of blocking due to a feedback mechanism between the planetary and subplanetary scales. If this hypothesis is true and if the feedback process begins with the subplanetary-scale waves, then these results imply that intrinsic errors would not allow for accurate simulations of blocking beyond 6 days. Specifically, the successful development of a planetary-scale blocking pattern may be impossible after just a few days if forced by unpredictable synoptic-scale eddies. Results from the following sections suggest that an alternate or expanded hypothesis should be considered.

2) PLANETARY SCALES

The component of the SS related to the phase error of the planetary waves is presented in Fig. 8a. To indicate phase error growth, the coefficient of determination (appendix A) between the perturbed and control height anomalies is subtracted from unity. As indicated by the solid horizontal line, phase accuracy is lost when the anomaly correlation reaches a value of 0.5 (Murphy and Epstein 1989). Over the first 8 days of the simulations, the average phasing error for the 8-day lead ensemble is larger than the other groups. This result is consistent with the false alarm signature noted in Fig. 4. However, at the end of this 8-day period, the growth of phase errors in the 8-day lead ensemble slows concurrent with the onset of the control block and an increase in the frequency of perturbed blocks (Fig. 4). This result suggests that the perturbed simulations tend to form blocks in nearly the same location as the control event. Phase accuracy is lost in the 0-day lead ensembles by day 11, the time at which some of the control blocking events begin to weaken and/or break down (Table 1). Phase accuracy is extended to about 12 days for the zonal and 2-day lead ensembles. Finally, the limit of phase accuracy reaches about 13.5 and 15 days for the 4- and 8-day lead ensembles, respectively. This result is remarkable, especially in view of the fact that the 2-, 4-, and 8-day lead ensembles cover the period of time when the control simulations make the transition to blocking. In spite of intrinsic errors, phase correlation of the planetary-scale waves is maintained in the ensemble simulations through the blocking transition, and extends well beyond 12 days for 4- and 8-day lead times.

The component of the SS related to the amplitude of the planetary waves (appendix A) is presented in Fig.
8b. On average, the amplitude error in the zonal ensemble grows rapidly after about 4 days. Conversely, the 0-, 2-, and 4-day lead blocking ensembles maintain an accuracy level near 0.15 for about 9 days. The exception is the 8-day lead ensemble, whose larger amplitude errors are related to the false alarms. Beyond about day 9, the mean amplitude error of the 8-day lead ensemble recovers and then grows at about the same rate as the 4-day lead group. Meanwhile, the amplitude errors of the 0- and 2-day lead ensembles grow quite rapidly after day 9. After making the transition to blocking, the 8- and 4-day lead ensembles have the smallest amplitude errors, and in fact, are still skillful at day 15.

The asterisks in Fig. 8b denote the fourth day after the onset of blocking in the control simulations, the time at which blocking frequency typically maximizes in the perturbed simulations and the control blocks are strongest (Fig. 4). The asterisks indicate that the mean wave amplitude error of the 0-, 2-, and 4-day lead ensembles is only approximately 0.1 when most of the simulations have developed a block. Significance tests (appendix B) reveal that the mean scores of the 0-, 2-, and 4-day lead ensembles do not differ significantly at the times denoted by asterisks. The mean amplitude error of the 8-day lead ensemble is larger than the other asterisked times, but its value is still relatively small (~0.2). Regardless of lead time, the wave amplitude errors of the simulations are relatively small at a time when the control blocks are strongest. Most of the blocking simulations not only make successful transitions to blocking, they also produce blocks with accurate amplitudes within a few days of the control simulations.

The total SS [Eq. (3)] for the planetary waves is presented in Fig. 8c. The scores reflect the features of phase and amplitude errors just discussed. Over the first 8 days of the simulations, the skill of the 8-day lead ensemble is lower than the other blocking ensembles. Again, this result is a consequence of phase and amplitude errors that are consistent with the blocking false alarm signature. The zonal ensemble saturates first at day 9, followed shortly by the blocking maintenance (0-day lead) ensemble. Thus, in the absence of model error, the predictability of blocking maintenance is comparable to that of the maintenance of the high-index regime. The onset of blocking ensembles reaches saturation at about 11.5, 13, and 14 days for the 2-, 4-, and 8-day lead groups, respectively, considerably later than the blocking maintenance cases. These results are supported at the 99% confidence interval as shown in Fig. B1.

3) FULL RESOLUTION SCORES

Mean skill scores for each of the ensembles has been examined in terms of synoptic-scale (Fig. 7) and planetary-scale (Fig. 8c) wavelengths. Mean skill scores using full-resolution (T42) height fields are shown in Fig.
Comparison of Fig. 8c and Fig. 9 reveals that only subtle differences exist between mean SS calculated for planetary-scale and full-resolution wavelengths. The greatest change is that the length of predictability for the 4-day (8-day) lead ensemble drops from around 13 days in the planetary waves to about 12 (11) days using full-resolution heights. Figure 9 demonstrates that the predictability characteristics of the ensembles are dominated by changes in the flow patterns that occur in the planetary waves.

d. Ensemble spread

Distributions of individual planetary-scale SS for each simulation are presented in Fig. 10. At any given time, the minimum and maximum values represent an estimate of the range of possible solutions that occur starting from slightly different, yet equally plausible, initial conditions. The scores are generally concentrated about the mean, as indicated by the shaded area within the first and third quartiles. Since the median score (solid line) is better than the mean score (dashed line), the distributions are negatively skewed toward less accurate simulations.

Each ensemble produces at least one simulation that is more than twice as accurate as climatology (i.e., SS > 0.5) at day 15. Conversely, each ensemble contains at least one simulation that saturates relative to climatology within about 6 days. So, while skill relative to...
climatology is maintained in the mean, a few simulations are superb while others are woeful. Toward the end of the simulations (day 12 or longer), only the 4- and 8-day lead ensembles have a majority of their members above climatology. Since a majority of these simulations (and the control) are blocked at this time, the data suggest that dispersion of the ensembles can be moderated in the presence of blocking. This result is intriguing, especially in view of the long lead time and regime transition that must be negotiated.

**e. Synoptic interpretations**

A representative sample of the daily 500-mb charts for blocking event number 3 (Table 1) is presented in Fig. 11 to aid with synoptic interpretation of the statistical results. The control case, C, represents days 8–13 of the 4-day lead simulations. The time period shown begins 4 days after the onset of the control block. This control case is compared with perturbed simulation numbers 5, 13, and 16 (P5, P13, P16) for that event. These examples are selected because they illustrate behaviors that are common among the ensembles. Specifically, P16 fails to amplify the planetary-scale waves but maintains accurate phasing with the control. On the other hand, P13 produces an accurate simulation of the planetary wave, but is not counted as a blocking event due to interruptions by passing synoptic-scale eddies. Finally, P5 represents an example of a simulation whose blocking magnitude exceeds that of the control.

All three of these simulations maintain remarkable phase coherence of the planetary-scale wave through day 12. With the 540-dam contour as a reference, note the longitudinal position of the trough and ridge axes for each simulation relative to the control. The evolution of these individual simulations is consistent with our statistical findings of the phase coherence of the perturbed simulations being quite accurate over the time period associated with the control event. Although still in phase with the control, the three examples demonstrate that a variety of solutions exist with regard to the amplitude of the planetary-scale waves. The simulations can remain locked in a nearly zonal pattern (P16), become overly strong (P5), or closely resemble the control event (P13). However, blocking frequencies (Fig. 4) indicate that most of the simulations do attain sufficient magnitude to meet the objective requirements for blocking, just like perturbed simulations P5 and P13. This observation is consistent with our statistical findings that the growth of the amplitude errors in the simulations tend to be diminished over time periods concurrent with the control block.

It was previously noted that the synoptic-scale waves become totally decorrelated by 6 days. Indeed, inspection of Fig. 11 reveals little correspondence between the eddies (wavenumbers > T10) in the control simulation and those in the perturbed simulations. In cases where the amplitude of the planetary-scale waves marginally support blocking, the superposition of synoptic-scale waves determines whether or not the pattern will be objectively identified as a blocking event. For example, on day 9 of P13, the superposition of the synoptic-scale wave (dashed contour) reduces the amplitude of the large-scale ridge. Consequently, the pattern does not meet the persistence requirements for a block because of this 1-day interruption.

Of course, not all of the perturbed simulations are as accurate as those shown in Fig. 11. In fact, some can be quite different, such as simulation number 5 (P5) of the 0-day lead ensemble for blocking event number 3 (Fig. 12). Note that simulation P5 in Fig. 12 is not the same as simulation P5 discussed in association with Fig. 11 that represented the fifth simulation of the 4-day lead ensemble. In this example (Fig. 12), the control block begins to decay after day 10. On day 10, planetary-scale waves for P5 exhibit a similar phase, but its amplitude is weaker. At the same time, a strong synoptic-scale wave is traversing through the Gulf of Alaska over the top of the planetary-scale ridge in P5. Note the lack of correspondence between this wave and the synoptic eddies in the control event. While the control block continues to decay, the synoptic-scale wave in P5 appears to play a role in the retrogression of the planetary-scale ridge as described by Berggren et al. (1949), Shutts (1983), Mullen (1987), and many others. By day 12, the planetary-scale height fields are nearly out of phase. Although simulation P5 is badly out of phase, it still qualifies as a blocking event. Perturbed simulations that evolve in a manner similar to P5 account for some of the largest dispersions depicted in Fig. 10.

**4. Concluding remarks**

In this paper, the impact of initial condition uncertainty on simulations of blocking onset, maintenance, and zonal flow was examined. Six events each of blocked and zonal flow were objectively identified within a 6-yr control simulation of the NCAR CCM1. For each of the identified events, 20 different initial conditions were specified by perturbing the mass and momentum fields to reflect realistic estimates of global analysis error. Case ensembles of perturbed simulations were generated for zonal flow and for blocking at lead times of 8, 4, 2, and 0 days (Fig. 2).

As lead time prior to the onset of the control block increases, the tendency to underestimate the frequency of blocking increases. Uncertainty in the timing of blocking onset also increases with lead time. These results are not surprising. As observed in Figs. 11 and 12, the synoptic-scale transient waves influence the development and maintenance of blocking ridges (e.g., Berggren et al. 1949; Shutts 1983; Mullen 1987; Hoskins and Sardeshmukh 1987; Nakamura and Wallace 1990), and can critically determine whether a day qualifies as a blocking event. However, these synoptic-scale waves become decorrelated among ensemble members that lack sufficient magnitude to meet the objective requirements for blocking.
What we find to be more intriguing, and very encouraging, is that the dispersion of the planetary-scale waves (wavenumbers $\leq T_{10}$) within the ensembles seems to be diminished over days associated the control block, even after the transition to blocking. This results in simulations that remain skillful relative to climatology for an extended period concurrent with the control blocking event. The 8-day lead simulations perform remarkably well, with a hit rate exceeding 60% at day 11 (Fig. 4) and the majority of simulations bettering climatology at day 14 (Fig. 10). By comparison, the accuracy of the zonal simulations drops steadily, reaching values associated with climatology within about 9 days.

It is interesting to note that each of the six blocking (Fig. 7), which may lead to enhanced uncertainty at larger scales.
(zonal) events discussed in this paper project strongly onto the positive (negative) phase of the model's Pacific–North American (PNA) teleconnectivity index (Wallace and Gutzler 1981; teleconnection maps are not shown). This implies that initial states in the CCM1 that project negatively are more sensitive to ICU, and thus are inherently less predictable than positive projections. This same behavior is also observed in the forecast models from several operational centers (e.g., Palmer 1988; Chen 1990; Tibaldi and Molteni 1990). Our perfect model results, in conjunction with those for the operational models, suggest that maintenance of the negative phase of the PNA pattern might remain more difficult to forecast than the onset or maintenance of the positive phase, even with future improvements in model formulations and data assimilation procedures.

Hoskins and Sardeshmukh (1987) have hypothesized that the midlatitude planetary flow must be predisposed for the onset of a blocking event. Colucci and Alberta (1996) have recently explored this hypothesis in a planetary-scale climatology of blocking. Our results seem
consistent with the idea that the model atmosphere may be preconditioned for the development of blocking. Even after a lead time of several days, long after the synoptic-scale waves become saturated relative to climatology, the simulations demonstrate a strong tendency to develop blocking events during the same period as the control block. While intrinsic error growth ultimately imposes limits on the predictability of blocking, the results of this study suggest that skillful, ensemble predictions of transitions to a blocked state are possible at long lead times if the model error is held to a minimum.

It is not known whether blocking simulations over different geographical regions (e.g., North Pacific vs North Atlantic), for different flow configurations (block vs cutoff lows, etc.), for different perturbation methods (e.g., Molteni et al. 1996), or from different models would behave similarly to the simulations examined in this paper. Future research along these paths seems warranted in order to understand thoroughly the predict-
ability characteristics of different flow regimes and regime transitions. Moreover, as greater computing resources and more realistic GCMs become available, a logical extension of this work is to perform an experiment in which the a priori selection of events is eliminated. By choosing thousands of independent (Madden 1979) initial flows at random from a very long (i.e., 100 years or longer) climatological simulation, differences in predictability characteristics could be examined for a variety of adequately populated flow regime and regime transition categories using, in part, the techniques outlined in this paper.

Acknowledgments. This paper is based on the master’s thesis of the first author and on a talk presented earlier (Nutter et al. 1996). The research was partially supported by National Science Foundation Grants ATM-9118898 and ATM-9319411. The authors thank the late Dr. A. Murphy for insights into error decomposition, and the anonymous reviewers whose constructive com-
ments helped improve the presentation of this manuscript.

**APPENDIX A**

**Climatological mse Score**

Following the work of Murphy and Epstein (1989), the mse in the 500-mb geopotential height field is used as a measure of accuracy. The model’s climatology is used as a standard of reference, defined by fitting a least squares parabola to the control data at each grid point. The advantage of using a climatological mse skill score resides in the fact that it may be decomposed into terms that address phase errors, amplitude errors (biases), and differences between sample and long-term climatologies (Murphy 1988, 1995; Murphy and Epstein 1989). This characteristic is especially useful for verification of blocking forecasts, where small phase shifts between the simulations and control height fields may lead to tremendous localized errors.

Verification of the simulations is performed within a region that is bounded to the northwest by a point near 76°N, 176°E and to the southeast by a point near 31°N, 120°W. These bounds correspond to an area approximately 45° latitude by 120° longitude, centered over western Canada. Within this region, there are \( N = 989 \) grid points from which the verification data sample is obtained. All calculations are weighted by cosine latitude.

For any given point in time, let \( p_n \) and \( c_n \) represent anomalies from the model’s climatological geopotential height fields at the \( n \)th grid point for the perturbed and control simulations, respectively. The spatial average of these anomalies over the verification area is denoted by angle brackets as

\[
\langle p \rangle = \frac{1}{N} \sum_{n=1}^{N} p_n \quad \text{and} \quad \langle c \rangle = \frac{1}{N} \sum_{n=1}^{N} c_n. \tag{A1}
\]

Standard deviations of the anomalies over the verification region are denoted as

\[
s_p = \left[ \frac{1}{N} \sum_{n=1}^{N} (p_n - \langle p \rangle)^2 \right]^{1/2}
\quad \text{and}
\]

\[
s_c = \left[ \frac{1}{N} \sum_{n=1}^{N} (c_n - \langle c \rangle)^2 \right]^{1/2}. \tag{A2}
\]

The sample covariance of \( p \) and \( c \) is

\[
s_{p,c} = \frac{1}{N} \sum_{n=1}^{N} (p_n - \langle p \rangle)(c_n - \langle c \rangle). \tag{A3}
\]

Finally, the product-moment anomaly correlation coefficient is written:

\[
r_{p,c} = \frac{s_{p,c}}{s_p s_c}. \tag{A4}
\]

Following the work of Murphy [1988, his Eq. (14)] the climatological mse skill score is decomposed as follows:

\[
SS = \left[ \frac{r_{p,c}^2 - [s_{p,c}^2 - (s_p s_c)]^2 - [(\langle p \rangle - \langle c \rangle)/s_p]^2 + (\langle c \rangle/s_c)^2}{1 + (\langle c \rangle/s_c)^2} \right]. \tag{A5}
\]

Each of the components in (A5) represent nondimensional measures of the contributions from different kinds of errors. The coefficient of determination, \( r_{p,c}^2 \), measures the correspondence in sign (phase) between the perturbed and control anomalies. Since \( r_{p,c}^2 \) contributes positively to overall skill, it may be considered a measure of potential skill in the absence of any biases. Biases in a forecast are measured by the conditional and unconditional biases, each of which contribute negatively to the overall skill. The conditional bias term, \( [s_{p,c}^2 - (s_p s_c)]^2 \), may be written \( [(s_p/s_c)(b - 1)]^2 \), where \( b \) is the slope of the regression line of \( p \) on \( c \). As such, the conditional bias statistically describes the relationship between the forecast and the average observation given that forecast. Thus, it is a measure of the bias that accounts for any lack of linear association between the simulations and control. The nondimensional measure of unconditional bias, \( [(\langle p \rangle - \langle c \rangle)/s_p]^2 \), is by definition 0 in a perfect model (if calculated over the entire domain). Although the verification area for this study is merely a subsection of the entire model domain, it was confirmed that the unconditional bias for this region remains negligible (i.e., an order of magnitude smaller than the leading terms) throughout the experiment. The remaining term in (A5), \( (\langle c \rangle/s_c) \), depends only on the height anomalies in the model control data. It contributes positively to the overall skill score, but has a relatively small value since the spatially averaged control anomaly is small compared to its standard deviation. In the current study, these nondimensional measures are used to interpret the initial condition error growth within the ensembles. For further discussion on the interpretation of these skill components, refer to Murphy (1988, 1995), Murphy and Epstein (1989), Stewart (1990), and Wilks (1995).

**APPENDIX B**

**Statistical Significance Testing**

The experimental design of this study prohibits the use of traditional statistical significance tests (e.g., Student’s t-test) because the premise of independent samples is violated. In particular, simulations within each ensemble are started by applying relatively small perturbations to the same basic state and the length of time between the 8- and 0-day lead ensembles is less than the 10 days required for independent sampling.

The method used in this study is commonly referred to as data rerandomization, bootstrapping, or Monte Carlo significance testing (Wilks 1995; Thiebaux 1994; Livezey and Chen 1983). Using the available data, an
empirical probability density function is created, from which statistical inferences are made. Since the empirical distribution is derived from the data itself, no assumptions are required and any appropriate statistic may be tested.

Before discussing the resampling procedure, an appropriate statistic must be chosen. For this study, the null hypothesis (H₀) that ensemble mean skill measures are equal at the 99% (α = 0.01) confidence level is tested. Analysis of variance techniques determine whether H₀ should be rejected in favor of the alternative—that the means are not all equal. But if H₀ were rejected using analysis of variance, we would not know which sample(s) caused the inequality. To retain this information, pairwise comparisons are made between the M sample means using Student’s t-test. Unfortunately, an experiment-wise error rate is introduced when making pairwise comparisons (Walpole and Myers 1989). The experiment-wise error rate is the probability of falsely rejecting H₀ in at least one of the pairs of means being considered. Given m comparisons, the probability of making this error is 1 − (1 − α)m. With five samples being compared m = 10 ways, the experiment-wise error rate for this study is approximately 0.1. Therefore, one of the 10 pairs of sample means will likely be considered different when indeed they are not.

To compensate for the experiment-wise error rate in pairwise comparisons, Duncan’s multiple range procedure is applied using Student’s t-test (Walpole and Myers 1989). When a set of Z means are ranked in ascending order, the range of any subset of those means is y − x + 1, where x < y ≤ M and x ≥ 1. The range for any given subset therefore has values between 2 and Z, inclusive. For a given range of sample means, the least significant difference (LSD) is the smallest value for which the difference in sample means is considered significantly different. Normally, the LSD is determined using degrees of freedom tables (Walpole and Myers 1989). However, since degrees of freedom are not easily estimated in this study, the LSD values are computed using the rerandomization procedure.

The procedure for significance testing via rerandomization is as follows:

1) Pool together all of the available data samples.
2) Select at random, without replacement, Z new samples from the combined pool.
3) Compute the mean of each sample and rank in ascending order (μ₁, . . . , μₙ).
4) For every possible pair of samples, compute the LSD, where LSD = (μᵢ − μⱼ)/√(s²/N). The LSD is a general form of the pooled t statistic, where μ is the sample mean, N is the number of members in the sample, and s² is the error mean square for the analysis of variance in a one-way classification. Store this value of the LSD in categorical range bin number y − x + 1.
5) Repeat steps one through four 1000 times, creating empirically derived probability density functions for each categorical range.
6) Compare pairwise differences between actual sample means with the LSD falling at the 99th percentile of the empirically derived PDF. If the actual difference in means was larger than the derived LSD, the differences are considered significantly different.

To simplify the presentation of results, the significance tests described above are applied to 4-day averaged skill scores. That is, within each simulation the scores are averaged over days 1–4, 5–8, and 9–12. The resampling procedure is then applied to determine if the ensemble means over these 4-day periods are significantly different. The results of the tests are presented in Fig. B1. For each of the skill score components, the 4-day averaged ensemble means are ranked in order from smallest to largest. Wherever a line connects the ensemble identifiers (Z, 0, 2, 4, or 8), the mean skill scores between those ensembles are not considered significantly different at the 99% confidence level if connected by a horizontal line.

### REFERENCES


Thompson, P. D., 1957: Uncertainty of initial state as a factor in the predictability of large scale atmospheric flow patterns. Tellus, 9, 275–295.