

Field Significance Revisited: Spatial Bias Errors in Forecasts as Applied to the Eta Model

KIMBERLY L. ELMORE,* MICHAEL E. BALDWIN,*⁺ AND DAVID M. SCHULTZ*

Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, Oklahoma

(Manuscript received 5 April 2005, in final form 1 June 2005)

ABSTRACT

The spatial structure of bias errors in numerical model output is valuable to both model developers and operational forecasters, especially if the field containing the structure itself has statistical significance in the face of naturally occurring spatial correlation. A semiparametric Monte Carlo method, along with a moving blocks bootstrap method is used to determine the field significance of spatial bias errors within spatially correlated error fields. This process can be completely automated, making it an attractive addition to the verification tools already in use. The process demonstrated here results in statistically significant spatial bias error fields at any arbitrary significance level.

To demonstrate the technique, 0000 and 1200 UTC runs of the operational Eta Model and the operational Eta Model using the Kain–Fritsch convective parameterization scheme are examined. The resulting fields for forecast errors for geopotential heights and winds at 850, 700, 500, and 250 hPa over a period of 14 months (26 January 2001–31 March 2002) are examined and compared using the verifying initial analysis. Specific examples are shown, and some plausible causes for the resulting significant bias errors are proposed.

1. Introduction

Forecast systems, in particular numerical forecast models, are always wrong. “How wrong are they?” is the question forecasters face daily, sometimes hourly. Sometimes the errors, while never exactly zero, are trivially small. Other times, the errors can be large and have serious adverse effects on the resulting forecast. Understanding and interpreting errors in model guidance is why human forecasters are an indispensable part of the forecasting process (e.g., Brooks et al. 1992). Fletcher’s (1956) prescience is still valid today: “The machine forecast always will be, on the average, really the *worst* product which the forecaster can put out—he can always do as well as the machine because it is his servant and he has the product at his disposal and he

always has the opportunity to let his brain (which the machine does not have) improve it.” One approach forecasters can use to improve upon a forecast is knowledge of the statistical structure and nature of spatial bias errors in the guidance and resulting forecasts.

Forecasters and modelers may use various approaches to understand errors inherent in numerical guidance. First, they might consider a comprehensive statistical measure. Perhaps the least useful for this application, but one easily implemented, is root-mean-square error (rmse) over the model domain as a function of forecast lead time. This is a poor choice for two primary reasons: rmse provides no insight into the spatial error distribution, nor does it yield any information about bias in the forecast.

Second, modelers and forecasters may try to understand how errors occur under similar circumstances through the use of analogs. Phenomenologically driven studies strive to gain insight into the nature of model errors for particular phenomena. For example, numerous studies are available concerning cyclone tracks and development, with lee cyclogenesis receiving particular attention (e.g., Mullen and Smith 1993; Smith and Mullen 1993; Schultz and Doswell 2000). Other studies examine how well models can forecast surface trough

* Additional affiliation: NOAA/National Severe Storms Laboratory, Norman, Oklahoma.

⁺ Additional affiliation: NOAA/Storm Prediction Center, Norman, Oklahoma.

Corresponding author address: Dr. Kimberly L. Elmore, NSSL, 1313 Halley Circle, Norman, OK 73069.
E-mail: kim.elmore@noaa.gov

passages (Colle et al. 2001) and tropical storms (e.g., Powell and Abernethy 2001).

Third, modelers and forecasters may need to understand errors in certain geographical regimes. Much previous work on model verification, particularly meso-scale model verification, has centered on limited areas. Three examples include Monobianco and Nutter (1999) who examine the 29-km Eta Model performance over the Florida peninsula; White et al. (1999) who show quantitative verification fields for six different models over the western United States, demonstrating notable differences between each model's performance; and Mass et al. (2002) who examine the relationship between model performance and horizontal grid spacing. Some work has also been aimed at how well an ensemble predicts the occurrence of a particular flow regime over limited areas (e.g., Chessa and Lalauette 2001).

Finally, another method examines spatial bias errors. Examples include Caplan and White (1989), Livezey and Chen (1983), Colby (1998), White et al. (1999), and Colle et al. (2000, 2003a,b). While conceptually straightforward, results from works like these may be difficult to interpret unambiguously. In other words, if a particular mean error is observed at a particular grid point, is that error statistically significant given *both* the variance at the grid point, and especially the overall *spatial* error variance structure of the error field? Studies that examine statistically significant spatial bias errors over extended periods are scarce. This is the question that motivates us because the *physical* significance and *physical* basis of bias errors cannot be generally addressed by either forecasters or model developers until those errors are known to be *statistically* significant. Errors that are not statistically significant may, or may not, have a coherent underlying cause. Yet, errors that are statistically significant are reliable in the sense that they are likely due to some physical weakness or inaccuracy in the model system or possibly in the initial conditions. Note, however, that statistical significance does not guarantee the ability to unravel the underlying physical basis for errors. Neither does statistical significance invariably connote physical significance. Small, statistically significant errors may not imply a physically or practically significant model error. No statistical method can account for careless experimental design, and no statistical method will identify the physical process underlying the error generation. Decomposing the nature and source of such errors is addressed in Murphy (1995).

In this paper, we use standard, easily implemented methods to find statistically significant differences between any two fields. These fields need not be on a

regular grid, though the examples that follow are performed on such a grid. The only necessity is that the points making up the forecast field must be accompanied by collocated verification values. Hence, this method may be used for assessing statistically significant errors in forecasts, defined either as the difference between a forecast and verification field, or the difference between two error fields from two different forecasts. The method is based on previous work by Livezey and Chen (1983) and utilizes straightforward, common statistical techniques that can be easily implemented on modern desktop computers. No special or proprietary software is needed to implement these techniques and they can be run automatically. The remainder of the paper is organized as follows. Section 2 discusses the technique applied to obtain statistically significant error fields. Section 3 describes the data used to demonstrate the analysis method. Section 4 provides a demonstration of the technique, and section 5 concludes with a discussion of the results.

2. Analysis technique

The analysis technique in this paper aims at establishing field significance, as described and defined by Livezey and Chen (1983). There are two separate significance levels that must be considered: local significance and field significance. Local significance tests statistical significance at individual grid points. A moving blocks bootstrap (Efron and Tibshirani 1993; Davison and Hinkley 1997; Wilks 1997) is used to create 95% confidence intervals ($\alpha_p = 0.05$) around the mean bias errors at each grid point. A moving blocks bootstrap is used because there is serial dependence in the errors at each grid point (the typical lag-1 autocorrelation in these data is about 0.2). A moving blocks bootstrap differs from a regular bootstrap in that the data are resampled in contiguous blocks, rather than by individual values (Fig. 1). This technique helps preserve the autoregressive structure within the data. Because moving blocks bootstrap resampling invariably results in some whitening of the time series, more sophisticated methods for postwhitening must be employed if serial dependence is a serious concern (Davison and Hinkley 1997). If the lag-1 autocorrelation at grid points that display the strongest autocorrelation is used as a benchmark, a block size of 7 seems reasonable for these data and this application (Fig. 2).

If the resulting confidence interval about the mean error does not contain zero, then the individual grid point is considered to possess statistically significant bias. If the confidence interval contains zero, the mean error at that grid point is not significantly different from

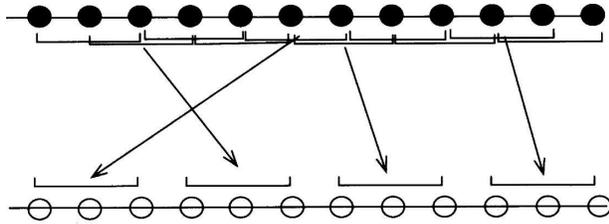


FIG. 1. A schematic diagram of the moving blocks bootstrap for time series data. The black dots are the original time series. A bootstrap resample of the time series (empty circles) is generated by choosing a block length (“3” in this example) and sampling with replacement from all possible contiguous blocks. Note that the block size need not be a multiple of the time series length, and that the set of blocks used for any particular resampling need not start at the first position (from Efron and Tibshirani 1993, used with permission).

zero. Hence, the result of this test is binomial at each grid point. While this process is similar in concept to the t test, it is free from parametric assumptions.

Field significance is not quite so straightforward and depends on spatial correlation. Spatial correlation describes how variations at one grid point are reflected at other grid points, due to physical processes covering areas larger than the grid spacing. Spatial correlation is present in nearly all meteorological fields. Were this not the case, plotted fields would possess no spatial coherence and would look like noise. The more structure a field possesses, the noisier it looks; the less structure it possesses, the smoother it looks. Fields rich in structure, but lacking in repeating patterns, tend to lack

spatial correlation. The amount of structure in a field may be likened to how many *independent* modes are contained in the field (Livezey and Chen 1983; Wang and Shen 1999).

Livezey and Chen (1983) appeal to the binomial distribution to develop a Monte Carlo method to determine what proportion of the grid points must yield statistically significant test statistic results for overall field significance of the test statistic at some α_f significance level (here, $\alpha_f = 0.05$, and the result at each grid point is binomial, with $\alpha = \alpha_p$, and number of trials $n =$ number of grid points). Their technique is also called the B (for binomial or Bernoulli) method in Wang and Shen (1999), who show that it is more accurate than either the χ^2 , the Z (which depends on the Fisher Z transformation), or the S (which assumes that the ratio of the mean variance over the variance of the mean yields the spatial degrees of freedom) methods. Wang and Shen (1999) also show that about 3000 Monte Carlo trials are needed for a standard error of about 10% in the spatial degrees-of-freedom estimate (or, alternatively, the estimate of the proportion of the grid occupied by significant errors). Hence, 3000 Monte Carlo trials are used here.

Briefly, the Monte Carlo process under the B method proceeds as follows: 1) generate a series of random numbers whose length equals the length of the error time series.¹ For example, if there are 100 days of data, then this series contains 100 elements. 2) Compute the correlation between this series and the time series of errors at each grid point; there are as many correlation values as there are grid points. 3) Determine what proportion of these correlation values is statistically significant. After all trials are complete, determine the $(1 - \alpha_p)$ quantile of the resulting distribution. If the proportion of grid points with significant bias exceeds this threshold, the spatial bias has field significance. Thus, the B method returns the proportion of grid points that could have significant errors at α_p purely by chance at some field-significance level, α_f . If the proportion of grid points with significant errors falls within the upper α_f tail of the Monte Carlo distribution, the errors possess field significance at α_f . To compare the errors between two different models requires only the difference between two error fields; the rest of the process remains identical.

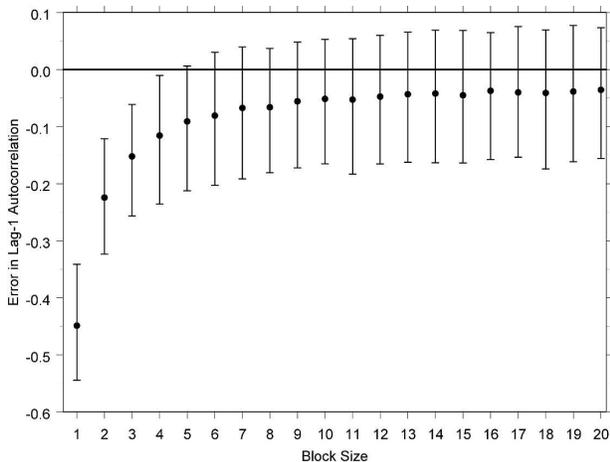


FIG. 2. The error in the block-bootstrap lag-1 autocorrelation estimate as a function of block size. The error bars show the 95% confidence interval for the lag-1 autocorrelation error. The horizontal reference line shows the zero error for the given series. Note that the mean underestimate of the lag-1 autocorrelation remains nearly constant as block size increases beyond about 7.

¹ Note that the random number series need not be normally distributed, though random numbers drawn from an $N(0, 1)$ distribution are used here. If \mathbf{U} is the vector of random numbers and \mathbf{X} is the time series at a grid point, then all that is required is that $E(\mathbf{X}\mathbf{U}) = E(\mathbf{U})E(\mathbf{X})$. In many cases, uniform random numbers are more stable, cheaper, and easier to generate.

Schematically, the steps of performing the analysis are as follows:

- 1) Grid point significance:
 - (a) Generate a matrix of forecast errors for each day, at each grid point. Thus, there will be as many columns of data as there are days, and as many rows as there are grid points.
 - (b) At each grid point, estimate the distribution of the mean forecast bias using a moving blocks bootstrap. For these data, a block size of about 7 works well.
 - (c) Compute mean error and confidence limits based on the desired α level at each grid point.
 - (d) Using the confidence limits computed in (c), determine what proportion of grid points contains significant bias errors, given by the number of grid points with bias errors significant at α_p divided by the total number of grid points.
- 2) Spatial mode computation:
 - (a) Use the Monte Carlo method developed in Livezey and Chen (1983) to estimate the distribution of the proportion of grid points that could contain significant bias purely by chance.
 - (b) Compare this to the proportion of grid points that possess significant bias errors. If the proportion of grid points with significant bias errors falls within the upper α_f tail of the Monte Carlo distribution, then the bias errors have field significance.

This approach is attractive because the entire process is automated through the moving blocks bootstrap and Monte Carlo approach; at no point in the analysis does the analyst have to actually know the spatial degrees of freedom. In addition, this approach is practically non-parametric, depending only upon resampling techniques and Monte Carlo simulation (the Monte Carlo simulation step uses as the null distribution the binomial distribution with as many independent trials as there are grid points and so is not strictly distribution-free). Of course, the analyst must still examine the resulting error fields and make subjective judgments about their reasonability.

3. Data

Selected output from the National Centers for Environmental Prediction (NCEP) operational Eta Model

(hereafter referred to simply as Eta; Black 1994) and a version of the Eta run locally at the National Severe Storms Laboratory (NSSL) called the EtaKF (Kain et al. 2001) are archived. The EtaKF model differs in three significant ways from the Eta: 1) it uses the Kain–Fritsch convective parameterization scheme (Kain and Fritsch 1990, 1993; Kain 2004), 2) a different shallow convective scheme is invoked (Baldwin et al. 2002), and 3) a fourth-order diffusion scheme, rather than the second-order scheme in the Eta (Kain et al. 2001) is used. Data are archived at the Storm Prediction Center (SPC) and NSSL.

Archival commenced on 26 Jan 2001 and the data were cut off for analysis on 31 Mar 2002. Because the Eta Model changed grid resolution during this period (discussed further in section 4), all data are nevertheless interpolated to the Automated Weather Integration and Processing System (AWIPS) 212 grid with 40-km horizontal grid spacing. Only geopotential, u , and v are archived for 850, 700, 500, and 250 hPa. To conserve storage space, only that part of the grid that fully encompasses the continental United States (CONUS), a fraction of southern Canada, and a fraction of northern Mexico is archived.

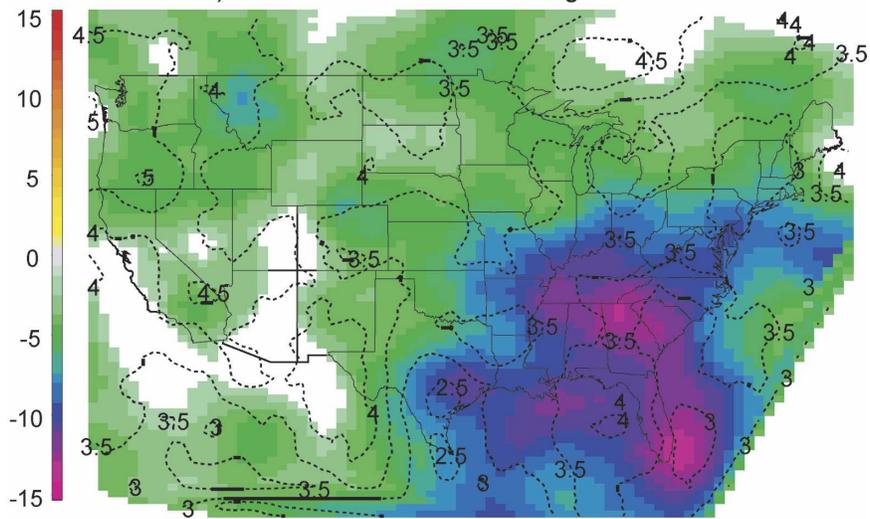
Data for this analysis consist of 24-h forecasts from the 1200 and 0000 UTC runs and their verifying analyses. Any arbitrary forecast lead time could be used, but the 24-h period is of particular interest to SPC forecasters for their day-2 outlook product. Over the 429-day period from 26 Jan 2001 to 31 Mar 2001, there are periods of missing data. Hence, for the Eta there are 406 0000 UTC forecasts and 380 1200 UTC forecasts with verifying analyses. For the EtaKF, there are 323 0000 UTC forecasts and 318 1200 UTC forecasts with verifying analyses. The Eta data contain 13 272 grid points, whereas the EtaKF contains 11 530 data points for geopotential and 11 552 data points for winds. Intercomparisons between the models utilize a common set of 11 530 grid points.

Using the Eta analysis for verification has drawbacks. The most serious concern is how much the 24-h forecast affects the verifying analysis. In locations where observations are scarce or nonexistent, the analysis will be highly biased by the previous forecast. In data-rich regions, such as the CONUS, the nature of the 4D Eta Data Assimilation System (EDAS; Black 1994), which

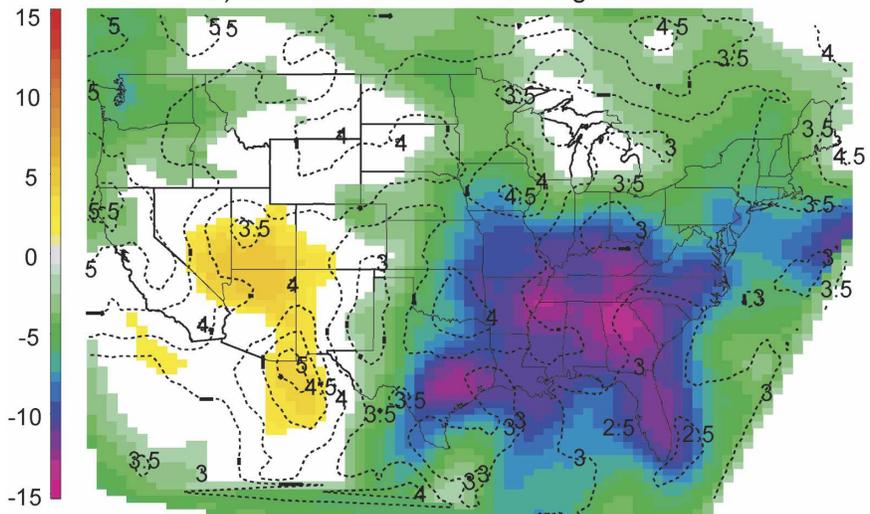
→

FIG. 3. Errors (m) for the (a) 0000 and (b) 1200 UTC 24-h 500-hPa forecasts, and (c) the difference between the 0000 and 1200 UTC errors (note different color scale limits). The observed coverage (minimum required coverage for 95% field significance) is 88.8% (11.4%) for the 0000 UTC errors, 78.1% (11.3%) for the 1200 UTC errors, and 45.0% (9.6%) for the difference between the two. Areas without shading are not significant at the 95% level, and dashed contours indicate the width of the 95% confidence interval (m), which is simply the upper value of the 95% confidence interval minus the lower value.

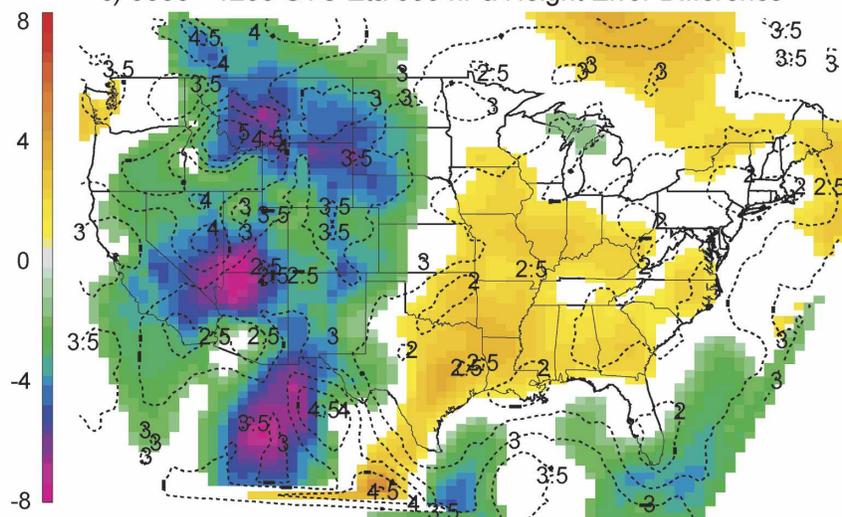
a) 0000 UTC Eta 500 hPa Height Errors



b) 1200 UTC Eta 500 hPa Height Errors



c) 0000 - 1200 UTC Eta 500 hPa Height Error Difference



includes a large amount of observational data, tends to mitigate this effect.

In all cases, the forecast values, the verification values, and the resulting errors are placed into matrices whose rows are the grid points and whose columns are the day. Once placed into matrices, forecast errors are defined as $\mathbf{E}_{24} = \mathbf{F}_{24} - \mathbf{O}$, where \mathbf{F}_{24} is the matrix of forecast values and \mathbf{O} is the matrix of verifying values. Thus, \mathbf{E}_{24} is a $13\,272 \times 406$ matrix for the 0000 UTC Eta errors, a $11\,530 \times 318$ matrix for the 1200 UTC EtaKF geopotential errors, etc. In this way, all statistics can be computed efficiently and simply. For example, the row means of \mathbf{E}_{24} represent the mean errors at each grid point and individual days, or strings of days (for a season, say). Selected rows may be extracted representing limited areas or “patches” for analysis.

4. Results

To show all verification statistics for all possible fields and all model differences is not practical in this venue. Hence, our examples are limited to those that illustrate certain points. In addition, we do not intend to dissect exhaustively Eta Model performance within this paper. We show particular examples of ways in which this technique might be used by developers and forecasters to understand physical processes behind model performance.

a. Model updates

Over the 14 months of data analyzed in this paper, the operational Eta Model underwent two major updates (<http://www.emc.ncep.noaa.gov/mmb/mmbpll/eta.log.para.html>). The first major update became operational with the 1200 UTC 24 July 2001 cycle (Rogers et al. 1999). This update included a modified three-dimensional variational data assimilation (3DVAR) analysis, used the 4-km NCEP National Precipitation Analysis (stage II) in the Eta data assimilation system, and made extensive modifications to the Eta Model land surface physics. The second major update increased horizontal and vertical resolution from 22 km/50 levels to 12 km/60 levels, introduced a new cloud microphysics scheme, and used an improved 3DVAR initialization. The second major update became operational with the 1200 UTC 27 November 2001 cycle. These two changes produce three subsets of data.

Are statistically significant changes in the spatial bias errors associated with any of these Eta updates? While a good question, the lack of parallel runs between the post and premodified model versions of two model ver-

sions poses a serious problem because any significant differences may be confounded by seasonal variability. If there are no statistically significant differences, model-based differences are unlikely.

An additional problem is that there are different numbers of days within each data subset. While generating bootstrap confidence intervals about the mean difference between different-sized datasets poses no problem, the Monte Carlo correlation step requires a matrix of error values at each time for each grid point. Further, to retain the serial correlation structure within the errors requires that all data in each matrix be consecutive, so data cannot be randomly extracted from the larger dataset.

An obvious way around this dilemma is to truncate the larger dataset so that there are as many days in it as in the smaller dataset. Unless no differences are found, or at least only marginally significant differences, exhaustively comparing each possible subset of consecutive days that can be extracted from the larger dataset to the smaller dataset may not be worth the effort. Not only is such an exercise computationally expensive, but the lack of parallel runs prevents any definitive results. Hence, the central contiguous period in the larger dataset is compared against the smaller dataset for only a few select fields (500-hPa height, 850–700-hPa thickness, and 700–500-hPa thickness).

The spatial bias errors between the different versions of the Eta are highly significant for both the 0000 and 1200 UTC cycles for all of the fields mentioned above; in the most marginal case, the threshold coverage for significance is exceeded by a factor of 2, and more often by a factor of 4. Even so, without parallel runs (as are always performed and available within the Environmental Modeling Center at NCEP), there is no good way to determine if this difference is due to seasonal changes or changes inherent in different model versions.

b. The 0000 versus 1200 UTC initializations and interseasonal differences

Another aspect of our analysis compares the 0000 and 1200 UTC runs of the Eta to test for a significant difference in the spatial bias between these two initializations. Because there are a different number of days in the 0000 and 1200 UTC datasets, the same technique used for the different innovations of the Eta Model must be used here, and in any other instance when the two datasets to be compared cover different periods.

As an example, height errors for the 24-h forecasts initialized at 0000 and 1200 UTC both possess field-

significant bias errors. While the bias errors are similar, they are not identical (Figs. 3a,b). The difference between these two fields is significantly different from zero for 54.8% of the grid points. The Monte Carlo test with 3000 trials requires significant results for only about 9.8% of the grid points for field significance. Hence, the difference between the 0000 and 1200 UTC 24-h height forecast errors is highly significant. The difference between the 850–700-hPa 0000 and 1200 UTC thickness errors (not shown) also possesses field significance, as 32.8% of the grid points are significantly different from zero, and the Monte Carlo test requires only 10.2%. Significant differences between 24-h forecasts that result from the 0000 and 1200 UTC initializations still exist even when the data are separated into seasons, where seasons are defined as DJF = winter, MAM = spring, JJA = summer, and SON = autumn (Table 1).

c. Spatial bias errors in thickness in the Eta and EtaKF models

To show the effect that different model formulations can have on the spatial bias errors, the thickness errors at different layers are evaluated and compared for the 0000 UTC Eta and EtaKF models for all available days using the methods described in section 3. Thickness

TABLE 1. Coverage (in %) of the difference between 24-h forecast errors resulting from the 0000 and 1200 UTC initializations of the NCEP Eta Model. The larger the difference between the required coverage and the observed coverage, the more significant the results.

Pressure level (hPa)	Observed coverage of significant difference ($\alpha_p = 0.05$)	Required coverage for field significance ($\alpha_f = 0.05$)
850, full period	55.1%	9.9%
700, full period	59.4%	10.2%
500, full period	45.0%	9.6%
250, full period	57.4%	9.9%
850, winter	28.9%	10.6%
700, winter	39.6%	10.9%
500, winter	33.6%	10.4%
250, winter	37.4%	11.2%
850, spring	46.7%	10.4%
700, spring	42.7%	10.2%
500, spring	29.7%	9.7%
250, spring	29.3%	10.2%
850, summer	23.7%	11.1%
700, summer	27.3%	10.9%
500, summer	27.3%	9.7%
250, summer	33.4%	10.0%
850, autumn	35.1%	11.4%
700, autumn	45.5%	11.1%
500, autumn	31.9%	10.7%
250, autumn	36.9%	10.5%

TABLE 2. Mean-layer height errors and equivalent U.S. Standard Atmosphere, 1976 mean-layer temperature errors.

Layer	Height error	Temperature error
850–700 hPa	1 gpm	0.17 K
700–500 hPa	1 gpm	0.10 K
500–250 hPa	1 gpm	0.05 K

errors represent mean-layer temperature errors and hence may act as a proxy for vertical temperature errors (Table 2). All three layers for both models possess field-significant spatial bias errors, and the differences between the spatial bias errors also possess unambiguous field significance (Table 3). Hence, not only are there significant errors within the models themselves, but the errors are significantly different between the models.

Thickness errors in the 0000 UTC Eta Model are negative over most of the domain in the 850–700- and 700–500-hPa layers (Figs. 4a,c), but positive over most of the domain in the 500–250-hPa layer (Fig. 4e). Although the errors are similar for the 0000 UTC EtaKF in the 500–250-hPa layer (cf. Figs. 4e,f; maximum values of +8.8 m for the Eta and +10.8 m for the EtaKF in similar regions), the differences between the Eta and EtaKF are substantial in the lower to midtroposphere, especially in the eastern United States (cf. Figs. 4a,b and 4c,d). In this case, the eastern United States is primarily associated with negative (cold) errors in the Eta and positive (warm) errors in the EtaKF (Figs. 4b,d). That the contours in the 850–700-hPa thickness field errors in the Eta parallel the coastline (Fig. 4a) suggests that the physical basis for these errors may be related to the supply of heat and/or moisture from the water and its redistribution to the lower troposphere.

TABLE 3. Similar to Table 1, but for thickness errors in both the Eta and EtaKF, along with thickness error differences between the Eta and EtaKF.

Difference between Eta and EtaKF thickness errors (by layer)	Observed coverage of significant difference ($\alpha_p = 0.05$)	Required coverage for field significance ($\alpha_f = 0.05$)
850–700 hPa	77.8%	9.4%
700–500 hPa	80.1%	10.0%
500–250 hPa	53.0%	10.5%
850–700 hPa (summer)	71.6%	9.3%
700–500 hPa (summer)	81.0%	9.0%
500–700 hPa (summer)	73.2%	10.0%
850–700 hPa (winter)	35.5%	9.6%
700–500 hPa (winter)	23.5%	10.0%
700–500 hPa (winter)	22.8%	11.4%

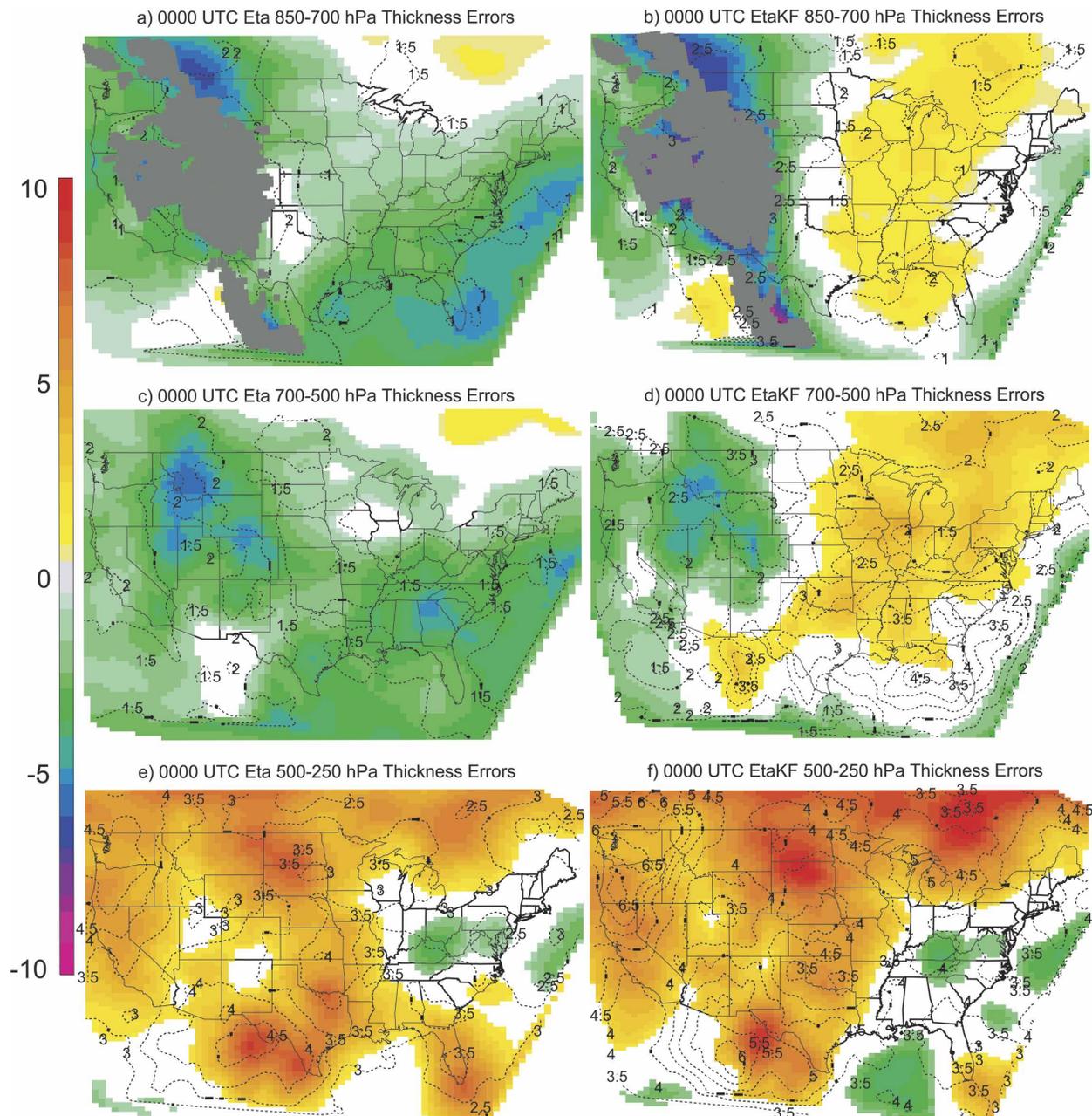


FIG. 4. Similar to Fig. 3, but for thickness errors for the 0000 UTC (left) operational Eta and (right) EtaKF significant at the 95% level for the (top) 850–700-, (middle) 700–500-, and (bottom) 500–250-hPa layers.

Seasonal differences also exist between the two formulations. Differences in the thickness errors in the 850–700-hPa and 700–500-hPa layers are more positive (warmer) in the Eta and EtaKF during the summer than in the winter (Figs. 5a–d). The 850–700-hPa thickness errors over the water over the eastern United States are nearly unchanged (Figs. 5a,c), which is evidence that the physical basis for these errors is unaffected by the seasonal cycle. It is certainly fair to inter-

pret Table 3 as showing an even higher confidence level for summer than for winter.

In the winter (Figs. 5c,d) over land, the thickness errors in the 850–700 hPa-layers were very similar in the Eta and EtaKF runs. Because the convective parameterization schemes are less active in winter than in summer, the summer differences must be strongly driven by the differences in convective parameterization schemes. Over the Gulf of Mexico and eastern

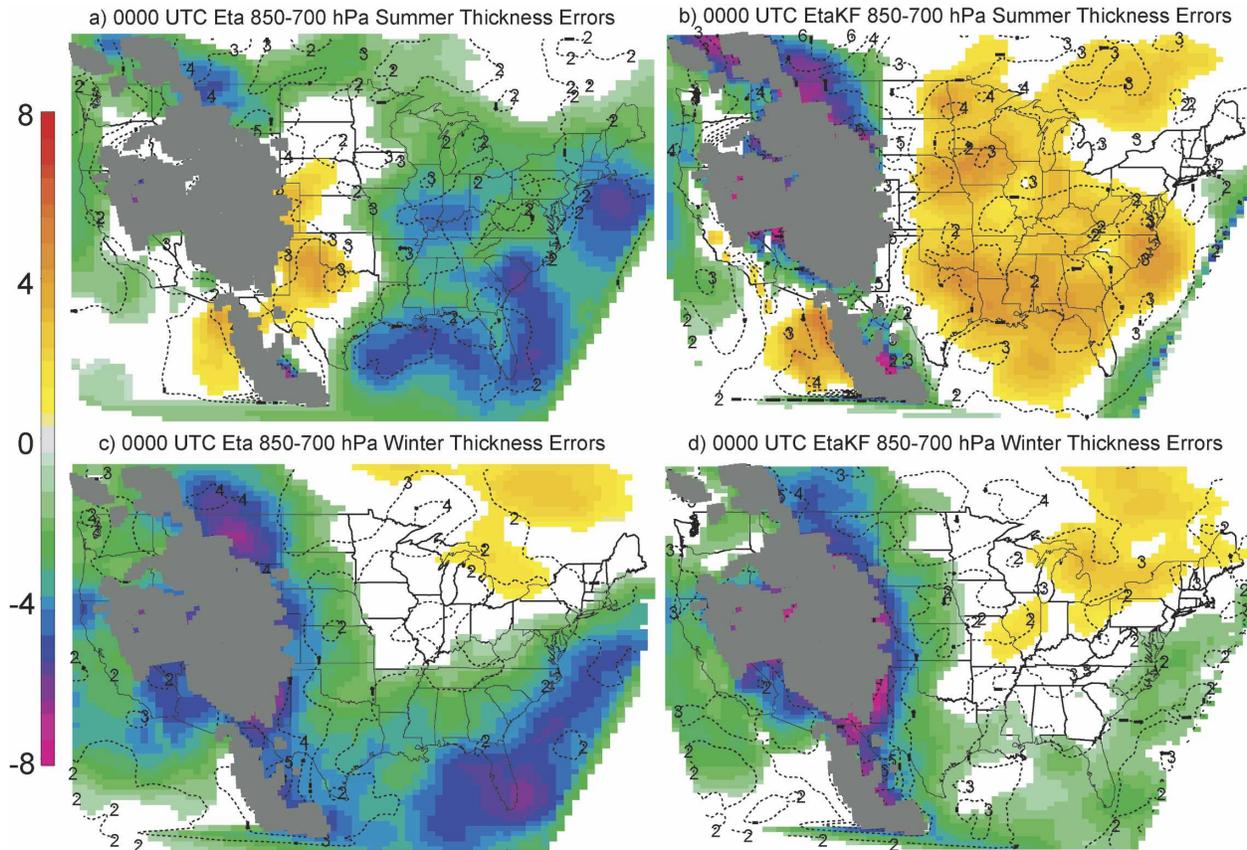


FIG. 5. Same as Fig. 4 but for 850–700-hPa (top) summer and (bottom) winter thickness errors.

United States, the thickness errors in the Eta are negative, while in the EtaKF over that same region thickness errors are positive. Nevertheless, the character of the thickness errors changes in the Eta over the western plains, where either positive errors or no significant mean errors are found. In that same region, the EtaKF also has positive errors or biases that are not statistically significant. In general, the Eta 850–700-hPa-thickness biases are the lowest (coolest) over the ocean water, and those biases tend to become more positive with increasing surface elevation.

A key difference between the Eta and EtaKF runs that may help to explain the disparity in the spatial structure of thickness biases is the parameterization of shallow convection. As discussed by Baldwin et al. (2002), the Eta uses the Betts–Miller–Janjić (BMJ; Betts 1986; Betts and Miller 1986; Janjić 1994) shallow convective scheme, which vertically mixes heat and moisture through the shallow cloud layer. In particular, the BMJ scheme transports heat downward from cloud top to cloud base and mixes moisture upward from cloud base to cloud top. The specific vertical layers affected by this mixing are determined by the position

of the shallow cloud in the vertical. Cloud base is assigned at the lifting condensation level determined from the most unstable layer in the lower part of the model atmosphere. Cloud top is defined as the model level within 200 hPa above cloud base where the relative humidity decreases the most with height. The EtaKF run also has a shallow convection component (Kain 2004), but differs in that CAPE is required in the cloud layer in order to produce active shallow convective mixing. Therefore, the overall magnitude and vertical extent of the mixing is considerably less than what is typically produced by the BMJ scheme in the Eta.

In regions where shallow convection is active, the Eta is inclined to be cooler than the EtaKF in layers affected by the upper part of the shallow convective process and warmer than the EtaKF in the lower part of the shallow cloud. Over the Gulf of Mexico and eastern United States, the Eta is significantly cooler than the EtaKF. Given the relatively low surface elevations in that region, the shallow cloud top is typically found near the 850–700-hPa layer, and the overall impact of BMJ shallow convection in that layer is to cool and lower the thickness values. On the other hand, in the

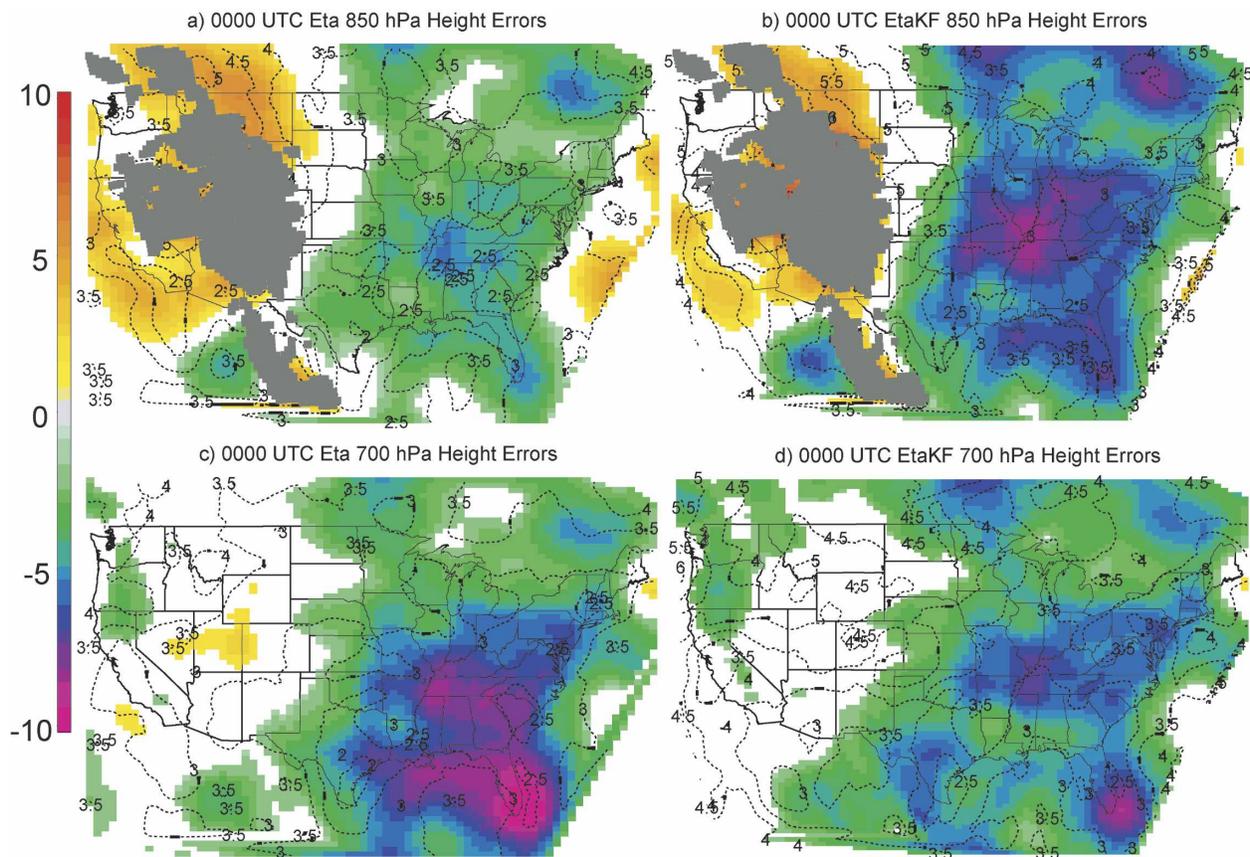


FIG. 6. Same as Fig. 5, but for (top) 850- and (bottom) 700-hPa geopotential height errors.

region of higher surface elevation and drier boundary layer conditions typically found across the western plains, the shallow cloud is likely to be positioned relatively high in the atmosphere. Warming associated with BMJ shallow convection near cloud base is speculated to be the primary cause for the positive 850–700-hPa-thickness bias found in the Eta over the western plains.

d. Spatial bias errors in height in the Eta and EtaKF formulations

Spatial bias errors in geopotential height are a function of different vertical mass and temperature distribution errors within the models. All levels of both the 0000 UTC Eta and EtaKF models possess bias errors in height that have field significance, but the difference in error structure between models is significantly different only at 500 hPa (not shown). Yet, the thickness errors are significant for all layers, a result that may not be intuitive if the analysis is not performed directly on thickness, instead of indirectly on heights alone.

For the 0000 UTC initialization, both models possess negative bias errors in 850-hPa height east of the Rocky

Mountains, but the EtaKF errors are much more negative (Figs. 6a–d). At 700 hPa, both models also possess negative height errors, but the Eta height errors are considerably more negative than the EtaKF. Because thickness is the 700-hPa height minus the 850-hPa height, the EtaKF thickness is considerably larger than the Eta thickness. The difference between the mean thickness errors between the two models is also significant. Hence, the difference in bias errors of thickness between the two models has field significance. Because this result may not be obvious by inspecting the height fields alone, computing and then testing the derived quantities themselves is required, instead of inferring how the derived quantities might respond by examining the parent variables.

e. Spatial bias errors in wind in the Eta Model

The spatial bias errors in the 700-hPa winds in the 0000 UTC Eta are shown in Fig. 7. The 700-hPa winds are too strong and too westerly in the northern portion of the domain and too weak in the southern part of the domain. Some specific locations have noticeable errors

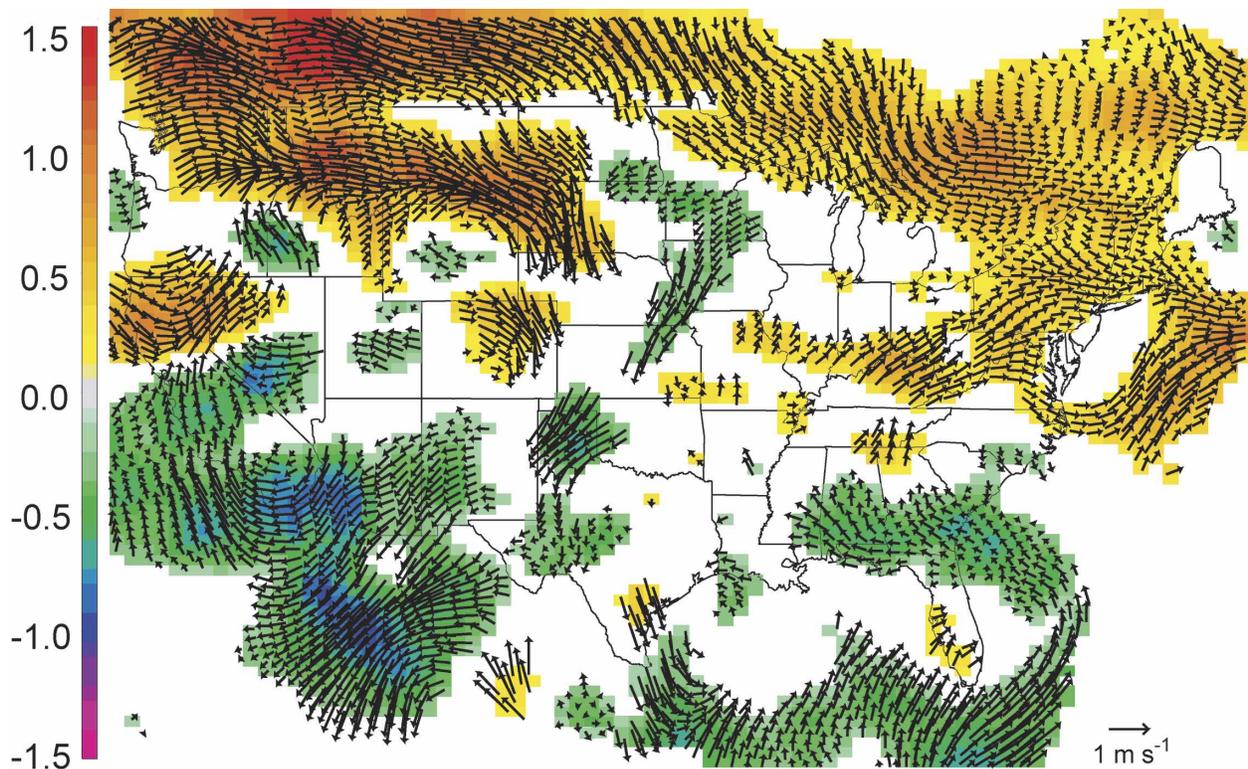


FIG. 7. Wind errors at 700 hPa for the 0000 UTC Eta based on total wind. Vectors represent error wind speed and direction while shaded areas represent the total wind error. Wind errors possess 95% field significance. A 1 m s^{-1} reference vector appears in the lower-right corner.

that may be easy to explain physically, though a conclusive explanation requires more data than are available for this work. For example, waves over the northwest United States and southwestern Canada suggest some aspects of flow over topography are not properly handled by the Eta Model. That these errors are persistent enough to appear in this analysis suggests a strong effect. Indeed, Gallus (2000) and Gallus and Klemp (2000) show that the step coordinate used by the Eta does not properly handle flow over orography. Another possibility is that the errors in stability associated with the thickness errors (Figs. 4a–f) lead to amplitude errors in the topographically forced waves.

5. Concluding discussion

In this paper, we reexamined the concept of field significance and illustratively applied these results to the NCEP Eta Model and an experimental version of the Eta with the Kain–Fritsch convective parameterization. We found that calculating field significance for model output can allow both model developers and forecasters to determine the statistical significance of spatial bias errors in a model. Thus, the approach de-

veloped by Livezey and Chen (1983) has been extended in a general way for spatial bias errors. In addition, these statistics can also be used to compare two model output fields to determine if they are significantly different from one another. Such an analysis could be used to evaluate (a) whether the bias errors of an updated model are statistically different (or improved) over a previous version (section 4a), (b) whether the spatial bias errors differ in a statistically significant way between seasons or between different model initialization times (e.g., 0000 versus 1200 UTC; section 4b), (c) whether two different models are statistically different (sections 4c and 4d), or (d) a general characterization of model error (section 4e).

A thorough understanding of spatial bias errors in numerical model output can also help model developers and forecasters shed light on whether systematic errors in the model are more likely related to errors in the initial conditions or errors in the model formulation (e.g., inadequate resolution or physical parameterizations).

Finally, although single numbers representing model errors (e.g., rms error) are a useful tool for model verification and evaluation, such measures lack spatial in-

formation. Such information can be useful to model developers because different regions are affected more by different physical processes. Spatial bias error information may be useful to forecasters because, for whatever reason, errors may have both seasonal and spatial variation. We reiterate here that these techniques need not be limited to forecast guidance data on a regular grid: the forecast data may be on *any* arbitrary grid subject to a verification value at each grid point. As shown in this paper, incorporating spatial bias errors into model verification studies provides a different, additional way to look at the model output. Forecasters can adjust (perhaps manually, or through Interactive Forecast Preparation System smart tools) forecasts based on characteristic spatial bias errors in the models and forecasts. Thus, spatial bias errors are a useful, although underutilized, approach for model and forecast verification studies.

Acknowledgments. We have benefited considerably from discussions with and comments from Dr. Brad Ferrier, Dr. David Stensrud, Dr. Michael Richman, and two anonymous reviewers of an earlier manuscript. Funding was provided by NOAA/OAR/NSSL under NOAA-OU Cooperative Agreement NA17RJ1227.

REFERENCES

- Baldwin, M. E., J. S. Kain, and M. P. Kay, 2002: Properties of the convection scheme in NCEP's Eta Model that affect forecast sounding interpretation. *Wea. Forecasting*, **17**, 1063–1079.
- Betts, A. K., 1986: A new convective adjustment scheme. Part I: Observational and theoretical basis. *Quart. J. Roy. Meteor. Soc.*, **112**, 677–691.
- , and M. Miller, 1986: A new convective adjustment scheme. Part II: Single column tests using GATE wave, BOMEX and arctic air-mass data sets. *Quart. J. Roy. Meteor. Soc.*, **112**, 693–709.
- Black, T. L., 1994: The new NMC mesoscale Eta Model: Description and forecast examples. *Wea. Forecasting*, **9**, 265–278.
- Brooks, H. E., C. A. Doswell III, and R. A. Maddox, 1992: On the use of mesoscale and cloud-scale models in operational forecasting. *Wea. Forecasting*, **7**, 352–362.
- Caplan, P. M., and G. H. White, 1989: Performance of the National Meteorological Center's medium-range model. *Wea. Forecasting*, **4**, 391–400.
- Chessa, P. A., and F. Lalauette, 2001: Verification of the ECMWF ensemble prediction system forecasts: A study of large-scale patterns. *Wea. Forecasting*, **16**, 611–619.
- Colby, F. P., Jr., 1998: A preliminary investigation of temperature errors in operational forecasting models. *Wea. Forecasting*, **13**, 187–205.
- Colle, B. A., C. F. Mass, and K. J. Westrick, 2000: MM5 precipitation verification over the Pacific Northwest during the 1997–99 cool seasons. *Wea. Forecasting*, **15**, 730–744.
- , —, and D. Ovens, 2001: Evaluation of the timing and strength of MM5 and Eta surface trough passages over the eastern Pacific. *Wea. Forecasting*, **16**, 553–572.
- , J. B. Olson, and J. S. Tongue, 2003a: Multiseason verification of the MM5. Part I: Comparison with the Eta Model over the central and eastern United States and impact of MM5 resolution. *Wea. Forecasting*, **18**, 431–457.
- , —, and —, 2003b: Multiseason verification of the MM5. Part II: Evaluation of high-resolution precipitation forecasts over the northeastern United States. *Wea. Forecasting*, **18**, 458–480.
- Davison, A. C., and D. V. Hinkley, 1997: *Bootstrap Methods and Their Application*. Cambridge University Press, 582 pp.
- Efron, B., and R. J. Tibshirani, 1993: *An Introduction to the Bootstrap*. Chapman and Hall, 436 pp.
- Fletcher, R. D., 1956: Two outstanding problems of modern meteorology. *Bull. Amer. Meteor. Soc.*, **37**, 473–476.
- Gallus, W. A., 2000: The impact of step orography on flow in the Eta Model: Two contrasting examples. *Wea. Forecasting*, **15**, 630–639.
- , and J. B. Klemp, 2000: Behavior of flow over step orography. *Mon. Wea. Rev.*, **128**, 1153–1164.
- Janjić, Z. I., 1994: The step-mountain Eta coordinate model: Further development of the convection, viscous sublayer, and turbulence closure schemes. *Mon. Wea. Rev.*, **122**, 927–945.
- Kain, J. S., 2004: The Kain–Fritsch convective parameterization: An update. *J. Appl. Meteor.*, **43**, 170–181.
- , and J. M. Fritsch, 1990: A one-dimensional entraining/detraining plume model and its application in convective parameterization. *J. Atmos. Sci.*, **47**, 2784–2802.
- , and —, 1993: Convective parameterization for mesoscale models: The Kain–Fritsch scheme. *The Representation of Cumulus Convection in Numerical Models*, Meteor. Monogr., No. 24, Amer. Meteor. Soc., 165–170.
- , M. B. Baldwin, and S. J. Weiss, 2001: Utilizing the Eta Model with two different convective parameterizations to predict convection initiation and evolution at the SPC. Preprints, *Ninth Conf. on Mesoscale Processes*, Ft. Lauderdale, FL, Amer. Meteor. Soc., 91–95.
- Livezey, R. E., and W. Y. Chen, 1983: Statistical field significance and its determination by Monte Carlo techniques. *Mon. Wea. Rev.*, **111**, 46–59.
- Mass, C. F., D. Ovens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bull. Amer. Meteor. Soc.*, **83**, 407–430.
- Mullen, S. L., and B. B. Smith, 1993: The dependence of short-range surface cyclone forecasts on the large-scale circulation: A preliminary assessment. *Wea. Forecasting*, **8**, 235–247.
- Murphy, A. H., 1995: A coherent method of stratification within a general framework for forecast verification. *Mon. Wea. Rev.*, **123**, 1582–1588.
- Monobianco, J., and P. A. Nutter, 1999: Evaluation of the 29-km Eta Model. Part II: Subjective verification over Florida. *Wea. Forecasting*, **14**, 18–37.
- Powell, M. D., and S. D. Aberson, 2001: Accuracy of United States tropical cyclone landfall forecasts in the Atlantic basin (1976–2000). *Bull. Amer. Meteor. Soc.*, **82**, 2749–2768.
- Rogers, E., D. Parrish, and G. DiMego, 1999: Changes to the NCEP operational Eta analysis. NWS Technical Procedures Bulletin 454, 25 pp. [Available online at <http://www.nws>.

- noaa.gov/om/tpb/454.pdf or Office of Meteorology, 1325 East-West Highway, Silver Spring, MD 20910.]
- Schultz, D. M., and C. A. Doswell III, 2000: Analyzing and forecasting Rocky Mountain lee cyclogenesis often associated with strong winds. *Wea. Forecasting*, **15**, 152–173.
- Smith, B. B., and S. L. Mullen, 1993: An evaluation of sea level cyclone forecasts produced by NMC's nested-grid model and global spectral model. *Wea. Forecasting*, **8**, 37–56.
- Wang, X., and S. S. Shen, 1999: Estimation of spatial degrees of freedom of a climate field. *J. Climate*, **12**, 1280–1291.
- White, B. G., J. Paegle, W. J. Steenburgh, J. D. Horel, R. T. Swanson, L. K. Cook, D. J. Onton, and J. C. Miles, 1999: Short-term forecast validation of six models. *Wea. Forecasting*, **14**, 84–108.
- Wilks, D. S., 1997: Resampling hypothesis tests for autocorrelated fields. *J. Climate*, **10**, 65–82.