Verification of Nowcasts from the WWRP Sydney 2000 Forecast Demonstration Project

ELIZABETH E. EBERT,* LAURENCE J. WILSON,+ BARBARA G. BROWN,# PERTTI NURMI,@
HAROLD E. BROOKS,& JOHN BALLY,* AND MATTHIAS JAENEKE**

*Bureau of Meteorology Research Centre, Melbourne, Victoria, Australia
+Recherche en Prévision Numérique, Dorval, Quebec, Canada
#National Center for Atmospheric Research, Boulder, Colorado
@Finnish Meteorological Institute, Helsinki, Finland
&NOAA/National Severe Storms Laboratory, Norman, Oklahoma
**Deutscher Wetterdienst, Offenbach, Germany

(Manuscript received 2 July 2002, in final form 16 December 2002)

ABSTRACT

The verification phase of the World Weather Research Programme (WWRP) Sydney 2000 Forecast Demonstration Project (FDP) was intended to measure the skill of the participating nowcast algorithms in predicting the location of convection, rainfall rate and occurrence, wind speed and direction, severe thunderstorm wind gusts, and hail location and size. An additional question of interest was whether forecasters could improve the quality of the nowcasts compared to the FDP products alone.

The nowcasts were verified using a variety of statistical techniques. Observational data came from radar reflectivity and rainfall analyses, a network of rain gauges, and human (spotter) observations. The verification results showed that the cell tracking algorithms predicted the location of the strongest cells with a mean error of about 15–30 km for a 1-h forecast, and were usually more accurate than an extrapolation (Lagrangian persistence) forecast. Mean location errors for the area tracking schemes were on the order of 20 km.

Almost all of the algorithms successfully predicted the frequency of rain throughout the forecast period, although most underestimated the frequency of high rain rates. The skill in predicting rain occurrence decreased very quickly into the forecast period. In particular, the algorithms could not predict the precise location of heavy rain beyond the first 10–20 min. Using radar analyses as verification, the algorithms' spatial forecasts were consistently more skillful than simple persistence. However, when verified against rain gauge observations at point locations, the algorithms had difficulty beating persistence, mainly due to differences in spatial and temporal resolution.

Only one algorithm attempted to forecast gust fronts. The results for a limited sample showed a mean absolute error of 7 km h^{-1} and mean bias of 3 km h^{-1} in the speed of the gust fronts during the FDP. The errors in sea-breeze front forecasts were half as large, with essentially no bias. Verification of the hail associated with the 3 November tornadic storm showed that the two algorithms that estimated hail size and occurrence successfully diagnosed the onset and cessation of the hail to within 30 min of the reported sightings. The time evolution of hail size was reasonably well captured by the algorithms, and the predicted mean and maximum hail diameters were consistent with the observations.

The Thunderstorm Interactive Forecast System (TIFS) allowed forecasters to modify the output of the cell tracking nowcasts, primarily using it to remove cells that were insignificant or diagnosed with incorrect motion. This manual filtering resulted in markedly reduced mean cell position errors when compared to the unfiltered forecasts. However, when forecasters attempted to adjust the storm tracks for a small number of well-defined intense cells, the position errors increased slightly, suggesting that in such cases the objective guidance is probably the best estimate of storm motion.

1. Introduction

The Sydney 2000 Forecast Demonstration Project (FDP) provided an unprecedented opportunity to test a large number of radar-based nowcast algorithms in an operational setting. Nine automated nowcast systems were run in parallel by a team of international researchers working alongside the severe weather forecasters in the Sydney office of the Australian Bureau of Meteorology (BoM). During the period September–November 2000 forecasters used the nowcast products from all nine algorithms to help them forecast rain, winds, and severe weather for periods of 1–3 h.

The subjective assessment of the nowcast products was quite positive. BoM forecasters agreed that the nowcast schemes indeed provided useful information for forecasting. In addition, they felt that being able to discuss the nowcast algorithms and products with the re-

*Corresponding author address: Dr. Elizabeth E. Ebert, Bureau of Meteorology Research Center, GPO Box 1289K, Melbourne VIC 3001, Australia.
E-mail: e.ebert@bom.gov.au

searchers improved their understanding of the physical processes that cause convection and severe weather. External users of the nowcasts (Sydney Olympic Committee, airlines, State Emergency Services, and Bridgeclimb) found the nowcasts beneficial for their smooth operations during the Olympic Games (Anderson-Berry et al. 2004, in this issue).

A more precise assessment of the nowcasts' skill requires objective verification against the observed weather. Algorithm verification is a vital component of the FDP and was built into the project early on. An international verification team, comprising the authors of this paper, was assembled to perform an independent nowcast verification using a variety of statistical and analytical techniques (Brown et al. 2001; Nurmi et al. 2001). They set out to answer the following questions:

1) Is it feasible to predict the location of convection with enough accuracy and skill to be useful?
2) What are the accuracy and skill of rainfall rate and occurrence forcasts as a function of lead time and accumulation period?
3) Is it feasible to predict wind speed and direction at points with enough accuracy and skill to be useful?
4) What is the accuracy of severe thunderstorm wind gust diagnoses and forecasts?
5) What is the accuracy of hail location and size detections and forecasts?
6) Do the forecasters improve the quality of the forecasts compared to the ''raw'' FDP products alone?

Initially it had been hoped that the team could provide real-time verification tools for use during the study period. Time constraints prevented this from happening, so the archived nowcast products were verified during the ensuing months.

There are several reasons for doing the nowcast verification. Most important, the verification gives algorithm developers specific information about the ability of their algorithms to forecast the types of weather situations that occurred in the Sydney 2000 FDP.[1] The results highlight situations in which the algorithms made good forecasts. Conversely, situations where the nowcasts were in error indicate weaknesses in the algorithms that require improvement. Forecasters, the main users of the nowcasts, use the verification results to get a feeling for the expected errors in the nowcasts, and especially to learn in which weather situations a given algorithm is particularly trustworthy or untrustworthy. Ongoing nowcast verification allows the systems to be monitored, and upgrades to be evaluated. Finally, it is intended that this nowcast verification, although completed after the FDP, will lead to the development of tools for improved real-time verification of nowcasts.

This paper starts by giving a brief description of the nowcast algorithms and their products, followed by a description of the data and methods used to verify the nowcasts. Selected verification results are shown for each algorithm. The paper concludes by returning to the set of six questions posed above.

2. Nowcast algorithms

The nine nowcast schemes run during the FDP are summarized in Table 1. More complete descriptions of the algorithms can be found in the references listed in the table, as well as in Keenan et al. (2003) and other papers in this special issue. All of the algorithms rely on radar reflectivity as their primary data source.

It should be emphasized that the algorithms were designed for differing purposes. The Canadian Radar Decision System (CARDS; Lapczak et al. 1999) and the Warning Decision Support System (WDSS; Eilts et al. 1996) were designed specifically for nowcasting severe weather, while Auto-nowcaster (ANC; Wilson et al. 1998), the Generating Advanced Nowcasts for Development in Operational Land-surface Flood Forecasts System (Gandolf; Pierce et al. 2000), and the Thunderstorm Identification, Tracking, Analysis, and Nowcasting System (TITAN; Dixon and Weiner 1993) are intended for thunderstorm nowcasting, and the Nowcasting and Initialisation for Modelling using Regional Observation Data Scheme (Nimrod; Golding 1998) and the Spectral Prognosis approach (S-PROG; Seed and Keenan 2001) are meant for rain nowcasting in general. The C-band polarimetric radar (C-Pol) approach is a hydrometeor classification scheme and does not produce forecasts. The Thunderstorm Interactive Forecasting System (TIFS, formerly known as Thunderbox; Bally 2004, in this issue) allows graphical interactive modification of any cell-based nowcast; in the FDP, TIFS used TITAN and WDSS nowcasts as its initial guess.

The algorithms predict a variety of meteorological quantities. The severe weather algorithms detect and predict tornadoes, hail, and downbursts, and all algorithms except WDSS and C-Pol predict rain intensity. The ANC algorithm is unique in nowcasting gust fronts and other boundary layer convergence lines. The spatial resolution of the algorithms ranges from pixel resolution (1 km in the FDP) to 5 km, and the temporal resolution ranges from 5 to 30 min. Forecasts extend from 10 min to 3 h.

Some algorithms are designed to make use of additional data sources such as numerical model forecasts, radiosonde soundings, satellite observations, and lightning data. There was not enough time before the FDP to reconcile the incompatibilities between the BoM's satellite and lightning data sources and the input streams of the algorithms, so unfortunately some of the algorithms were operating at less than their full potential (see Table 1). This must be taken into account when interpreting the verification results.

[1] Weather observed during the FDP included sea-breeze circulations in the Sydney basin, strong wind events, pre- and postfrontal precipitation, thunderstorm evolution and movement, and a supercell that spawned three tornadoes (Webb et al. 2001).

TABLE 1. Summary of the nine nowcast algorithms run in the Sydney 2000 FDP. See also Table 1 of Wilson et al. (2004, in this issue). The asterisks indicate algorithms that had reduced functionality compared to their original versions, due to some input data being unavailable. The full names of the algorithms and references are found in the text.

Algorithm	Origin	Purpose	Products	Spatial resolution	Temporal resolution	Comments
Auto-nowcaster*	NCAR	Thunderstorms	Boundary layer convergence lines, rain rate	Pixel	30- and 60-min forecasts, every 5 min	Expert system, storm interaction with convergence lines
CARDS	MSC	Severe weather	Hail size, mesocyclone, downburst detections; cell speed and direction, rain-rate forecasts	1 km	Every 5 min to 1.5 h	Modular design to run on variety of radars
C-Pol	BoM	Hydrometeor classification	Hydrometeor classification, rain rate	Pixel	Every 10 min	Fuzzy logic classification analyses
Gandolf	UKMO and University of Salford	Convective cells, rain nowcasting	Rain rate, rain accumulation	2 km	Every 10 min to 2 h	Object-oriented procedure with cell life cycle model
Nimrod*	UKMO	Short-period rain forecasts	Rain rate, rain accumulation	5 km	Every 30 min to 6 h	Radar-based nowcast tends toward model-based forecast with time
S-PROG	BoM	Rainfall nowcasting	Rain rate	1 km	Every 10 min to 1 h	Advection-based nowcast includes scale-dependent smoothing
TITAN	NCAR	Thunderstorms	Convective cell speed and direction	Pixel	30- and 60-min forecasts, every 5 or 10 min	Cell forecast based on weighted linear fit to historical storm track
WDSS*	NSSL	Severe weather	Tornado, hail size, downburst, lightning detection and probability; convective cell speed and direction	Pixel	Every 5 min to 1 h	Image processing, artificial intelligence, statistical methods to diagnose and forecast severe weather
TIFS	BoM	Interactive modification of cell forecasts	Convective cell speed and direction	Pixel	At discretion of forecaster	FDP version based on TITAN cell diagnoses

Quality control of the raw radar reflectivities is an important first step for the nowcast algorithms. Ground and sea clutter were serious sources of noise in the C-band Doppler radar located at Sydney Airport. Bright-band contamination and data dropouts were also issues. Although each algorithm has its own built-in quality control procedures, most of them were not optimally tuned for the Australian radar data (Donaldson et al. 2001). Incorrect association of clutter with precipitation was apparently a problem for some algorithms. As a result of these quality control issues, as well as spatial resolution differences, the analyses of the radar observations looked quite different from scheme to scheme, even though the same input data were used. An example of radar analyses for the tornadic storm of 3 November is shown in Fig. 1. Differences can be seen in the style and spatial resolution of the products, the magnitudes of the analyzed rain rates, and the misdiagnosis of sea clutter as light rainfall in some of the analyses.

3. Forecast variables and verification data

A set of meteorological variables to be verified was compiled prior to the start of the FDP, and included (a) convective cell location, (b) rainfall rate and occurrence, (c) wind speed and direction at point locations, (d) severe thunderstorm wind gusts at point locations, (e) hail location and size detections and forecasts, and (f) forecaster improvements to the quality of the automated nowcasts (Brown et al. 2001). After the FDP it became clear that the initial list would need to be revised, due to some forecasts and observations being unavailable. The final list of meteorological variables, and the observations that were used to verify them, is given in Table 2. Items c and d were removed, although some indirect assessment of wind nowcasts is possible with the verification of boundaries (convergence lines).

Verification data for all the systems came both from the radar analyses and from a network of rain gauges. The analyses were used to verify all those predictions that were defined spatially such as cell location, hail location, gust fronts, and other mesoscale convergence lines, and also were used to verify the predicted rain rates. Since we did not have available an analysis that was completely independent of all the nowcast systems, we felt the most consistent and fairest approach would be to verify each system against its own analysis.

Best estimates for the position of convective cells were determined by the cell tracking algorithms in TITAN and WDSS. The observed boundary layer convergence lines were manually inserted in real time by experts (J. Wilson 2001, personal communication), based on careful analysis of the Doppler radar wind velocities, and visually checked for accuracy against observations from the surface wind mesonet. Some uncertainty may arise in the detection of boundary layer convergence lines far from the radar, due to the increase in radar beam height with range; however, this is expected to have only a small impact on the results. To assess whether the ANC scheme correctly predicted the movement and spreading of gust fronts and sea-breeze fronts, both the position and length of nowcast convergence lines were verified against the manual analyses.

Rainfall was verified both against the radar rainfall analyses and against a network of 90 rain gauge observations shown in Fig. 2. The observations consist of accumulated precipitation over 5-min periods throughout the entire 3-month period of the experiment. Values are reported in increments of 0.5 mm. Comparison of the gauge time series with the radar rainfall analyses revealed no evidence of timing shifts between the event and the report in cases of light rain.

Verification of hail is difficult because hail observations are usually made by members of the public safety community or general public, and quality control for the observations is extremely difficult. Further, the absence of a report of severe weather does not necessarily mean that severe weather did not occur, only that it was not reported. The effect of this problem on verification of forecasts has been discussed by Brown et al. (1997). The FDP had few severe thunderstorm events that provided opportunities for collection of reports and for analysis of the severe weather algorithms. In fact, only one case (the storm that moved from Campbelltown to Castle Hill, producing a tornado near Parramatta on 3 November 2000) is suitable for analysis. Of the 20 reports of hail recorded in the forecaster's log in real time or on the following 2 days, only 11 contained useable size and location information; these are listed in Table 3. Clearly, this is not a complete description of the hail that fell out of the storm, but it is by far the best description available for any storm during the course of the experiment.

The verification results reported here were computed for a set of days and times during the 3-month operational period that were considered to have ''notable'' weather. These cases are listed in Table 4, categorized according to five different weather types: (a) most significant convection, (b) convective rain with reflectivity $\geq$40 dBZ, (c) convective rain with reflectivity $<$40 dBZ, (d) widespread stratiform rain, and (e) fair weather with sea-breeze circulations. The frontal types associated with the boundary layer convergence lines are also indicated in the table. The FDP contained 40 notable rain events and 16 cases of clear weather with sea-breeze fronts. The restriction of the verification dataset to interesting weather days had the effect of ''enriching'' the data, increasing the frequency of precipitation occurrence in the sample, and eliminating the majority of null cases (no precipitation forecast or observed).

4. Verification methods

A variety of verification methods were used to address the fundamental questions related to nowcast skill for convective cells, rainfall, convergence lines, and hail.

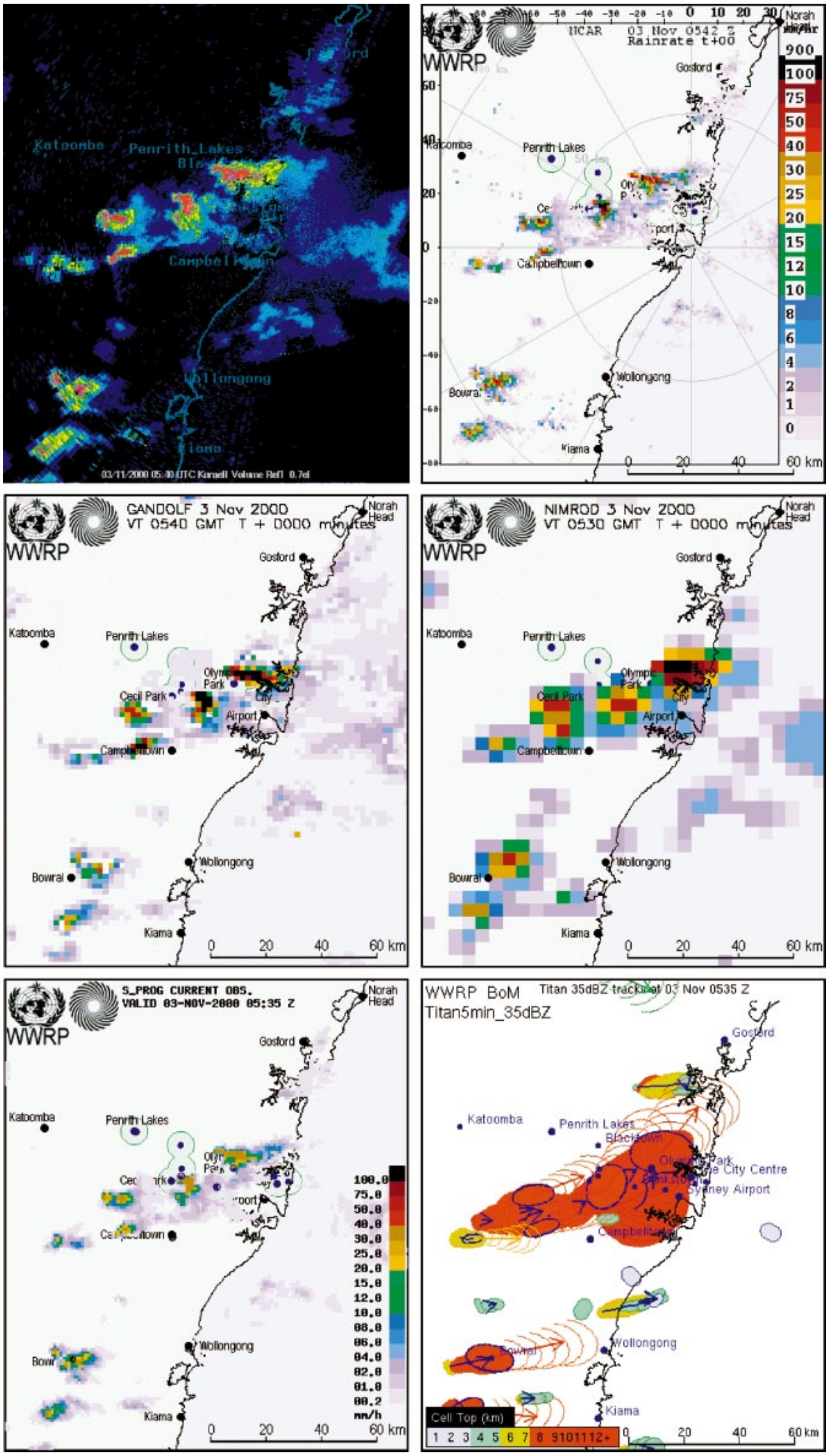

FIG. 1. Radar reflectivity and associated rainfall analyses from five nowcast algorithms at 0530–0540 UTC 3 Nov 2000.

TABLE 2. Meteorological variables predicted by the nowcast algorithms and verified using observational data.

Meteorological variables	Verifying observational data
Convective cell location	Radar reflectivity analyses
Rain rate and occurrence	Radar rainfall analyses*
Rain accumulation	Rain gauge observations
Boundary layer convergence line location and length	Manual analyses of Doppler radar wind velocities
Hail location and size	Human observations

* All algorithms used a climatological Z–R relationship, $Z = 200R^{1.6}$, to convert radar reflectivity to rain rate. No attempt was made to correct bias in the radar-rainfall estimates using rain gauge observations.

In some cases development of a new verification methodology was required. In this paper we present a subset of verification results that highlight the most important aspects of algorithm behavior. The statistics shown here are summary statistics, computed over the entire greater Sydney domain and over the events listed in Table 4. The results are stratified by forecast lead time and where possible by weather type.

Because of the numerous differences between algorithms we believe that it is inappropriate to directly intercompare their verification results. Therefore, each algorithm is evaluated individually in this paper. However, we do compare the verification results for each scheme to the results obtained for two alternate forecasts, namely persistence and extrapolation. The Eulerian persistence forecast (PERSIS) is defined as no change in the existing conditions; that is, the current weather is the forecast for the future weather. The simplest extrapolation forecast (EXTRAP), or Lagrangian persistence, is obtained by advecting the current radar analysis at a constant speed and direction over the entire domain, determined from prior mean storm motion using a simple correlation technique (see Seed and Keenan 2001 for details). Both of these forecasts are ‘‘unskilled,’’ in that no additional information or intelligence is used to predict the evolution of individual storms. If the nowcast algorithms perform better than PERSIS and EXTRAP, then they can be considered to add value to the forecast.

For cell and rain verification, only forecasts for which both the algorithm and EXTRAP forecasts were available were included to allow the two approaches to be fairly compared.

a. Cell location and velocity

The cell tracking feature of the TITAN and WDSS algorithms identifies convective cells as areas of contiguous reflectivity above a set threshold, and follows the centroid of each cell in consecutive radar images. The cell's speed and direction are determined from its past movement and used to forecast future cell position. Assuming that each cell is correctly and uniquely identified, it is simple to verify the forecast position against its observed position at the valid time of the forecast.

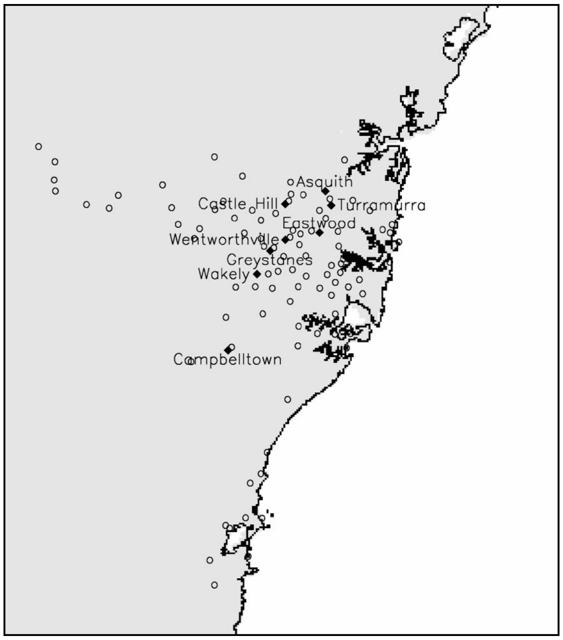

FIG. 2. Location of 90 rain gauges in the Sydney basin, denoted by open circles (o). The locations of hail sightings on 3 Nov 2000 are indicated by the filled diamonds.

For Cartesian nowcast algorithms without cell tracking, a modified version of Ebert and McBride's (2000) contiguous rain area (CRA) verification was used, with the 1 mm h^{-1} contour defining the CRA boundary. CRA verification uses pattern matching to estimate the position error of a forecast storm relative to its observed location, then verifies the ‘‘macro’’ properties of the storm, such as storm area, and the mean and maximum

TABLE 3. Hail reports with sizes associated with the 3 Nov 2000 severe thunderstorm (see Fig. 2 for locations of sightings). Comparison of the reported location to the observed track led us to question the timing of the Greystanes report. The location is approximately 15 km southwest of where the progression of reports would suggest that hail would be expected to be falling at 0600 UTC. It seems likely that the hail occurred at Greystanes approximately half an hour prior to 0600 UTC.

Time (UTC)	Location	Reported size	Equivalent diameter (cm)
0415	Campbelltown	Table tennis ball	3.2
0415	Campbelltown	10¢ coin (Australian)	2.4
0425	Campbelltown	20¢ coin (Australian)	2.9
0450	Northwest of Campbelltown	Golf ball	4.5
0505	Wakely	Golf ball	4.5
0541	Wentworthville	2 cm	2
0600	Castle Hill	2 cm	2
0600	Greystanes	Golf ball	4.5
0600	Eastwood	2 cm	2
0610	Turramurra	2 cm	2
0610	Asquith	1 cm	1

TABLE 4. Weather events observed during the FDP, Sep–Nov 2000 (J. Wilson 2001, personal communication). Frontal types associated with the boundary layer convergence line verifications are also indicated. SBF indicates a sea-breeze front, while the lowercase letters indicate the motion of other fronts (e.g., seF is a southeasterly front), usually gust fronts but occasionally a synoptic front. An asterisk indicates that TIFS nowcasts were made during this period.

Times and dates (UTC)	Frontal type	Comments
(a) Most significant convection		
1200–1900 25 Sep		Unexpected nighttime storms, downburst
2200 25 Sep–0230 26 Sep		Maximum reflectivities in upper 40s dB*Z*
0800-1300 26 Sep*	seF	Gust front and synoptic front collide
0200–1000 19 Oct*	neF	
0000–1530 3 Nov*	sF	Supercell with tornadoes
0000 30 Nov–1200 1 Dec*		Hail and flash flooding
(b) Convective rain with reflectivity $\geq$ 40 dB*Z*		
1900 9 Sep–1600 10 Sep		Over ocean
0300–0900 28 Sep	SBF	
0100–0800 29 Sep		
1930 13 Oct–0300 14 Oct	seF, wF	Mostly over ocean
0100–1000 19 Oct		50–55-dB*Z* storm south of Sydney
0200–0700 23 Oct	SBF	Intense slow-moving storm over Blue Mountains
0000–1400 26 Oct		Intensification upon moving offshore
2000 31 Oct–0400 1 Nov		Max reflectivities 45 dB*Z*
2200 4 Nov–1100 5 Nov*		
2200 5 Nov–1100 6 Nov	sF	
2200 6 Nov–1200 7 Nov		Max reflectivities 45 dB*Z*
0900–1600 20 Nov		
0100–1100 23 Nov		
0000–0600 24 Nov		
0000–2300 26 Nov		Severe storms reported
0000–9000 29 Nov		
(c) Convective rain with reflectivity $<$ 40 dB*Z*		
2100 8 Sep–0200 9 Sep		
0000–1000 17 Sep	SBF	
0000–0800 18 Sep	SBF	
1830 20 Sep–0900 21 Sep		
0300–1200 24 Sep		
1200 24 Sep–1000 25 Sep		
0000–0800 8 Oct	sF	
1000–2400 10 Oct		
1000 16 Oct–0200 17 Oct		
2000 1 Nov–0800 2 Nov	SBF	
(d) Widespread stratiform rain		
1100 27 Sep–0300 28 Sep	sF	
0800 8 Oct–1000 9 Oct		
1900 12 Oct–1610 13 Oct		
2200 17 Oct–0100 19 Oct		
1600 13 Nov–0900 15 Nov		
1900 15 Nov–1400 16 Nov		
1700–2400 16 Nov		
2000 17 Nov–1100 18 Nov		
(e) Fair weather with sea-breeze circulations		
2230 14 Sep	seF	
0120–0550 19 Sep	SBF	
2320 19 Sep–0330 20 Sep	seF	
0710–0830 22 Sep	stationary	
2320 29 Sep–0500 30 Sep	SBF	
0050–0720 3 Oct	SBF	
0050–0500 4 Oct	SBF	
0030–0630 5 Oct	SBF	
0350–0630 6 Oct	SBF	
0100–0410 10 Oct	SBF	
0440–0730 12 Oct	seF	
0120–0620 16 Oct	SBF	
0100–0520 20 Oct	SBF	
2230 26 Oct–0620 27 Oct	SBF	
1000–1110 27 Oct	sF	
0240–0740 17 Nov	SBF	

intensity. It also decomposes the total error into components due to displacement, volume, and spatial pattern errors.

A two-step pattern matching technique was implemented here. First, the forecast storm is translated over the observations until the overlap between the forecast and observed rain areas is maximized. Second, the pattern match is refined by searching in the immediate neighborhood for the location that minimizes the total squared error between the forecast and observations. This step matches the intense rain areas. To eliminate mismatches the pattern correlation between the two rain areas is required to be significantly greater than 0 at the 95% confidence level. This means that for very poor forecasts it is impossible to accurately estimate the position error using CRA verification. This might introduce a favorable bias in the verification results, although it is expected that the location errors determined for the identifiable cells will be representative of location errors in general.

b. Rain occurrence, rate, and amount

Rainfall was verified in a variety of ways. In all cases, it was necessary to first match the forecasts and observations in a consistent fashion. Then various verification statistics were computed from the sets of matched forecast–observation pairs. For verification of the spatial rain rates against analyses, the forecasts were simply matched pixel by pixel with the rain-rate analysis at the forecast valid time. Matching of the forecast rain-rate values with the gauge data was more complicated because of differences in the definitions of forecast and observation. The rain rates were matched spatially by selecting the rate forecast for the field pixel overlying each station. The predicted rain rates were considered to represent the average rain rate over the forecast output step of the model in each case (10 min for GANDOLF and S-PROG, and 30 min for ANC, Nimrod, TITAN, and CARDS), and were converted to accumulations over that period. The observations were available as accumulations over 5-min periods; these were matched by summing over the number of 5-min periods for each forecast step. The forecast was assumed to be centered over the observation period; that is, the forecast rain-rate value was interpreted to be in the center of the accumulation period for the purposes of time matching the forecasts and observations. Even with this optimal matching strategy, the differences in scale between spatially averaged radar snapshots and point accumulations at gauges lead to apparent errors, as will be seen in the next section. For verification with respect to observations, PERSIS is defined as the observed precipitation accumulation at the gauge at the initial forecast time.

The rain forecasts were verified both as continuous variables and as categorical variables. For verification as continuous variables, four measures were used. The linear bias, or mean error, is defined as the difference between the average forecast and average observation. Positive (negative) values indicate that the system is forecasting too much (too little) rain on average. The mean absolute error (MAE) is defined as the average magnitude of the difference between the forecast and observation. This gives an estimate of the expected magnitude of the error in the forecast. Third, a skill score based on the MAE was computed:

$$\mathrm{Skill}_{\mathrm{MAE}} = \frac{\mathrm{MAE}_{\mathrm{PERSIS}} - \mathrm{MAE}_{\mathrm{fcst}}}{\mathrm{MAE}_{\mathrm{PERSIS}}}.$$

This skill score measures the MAE of the forecast relative to the MAE of the persistence forecast, as defined above. It expresses the fractional improvement of the forecast over the standard forecast, and can range from $-\infty$ to 1. If the forecast is more accurate than the persistence forecast, the skill value will be positive, while negative values indicate that persistence gives a more accurate forecast, according to the MAE. The correlation coefficient between the forecast and observed rain measures how well the spatial or temporal pattern is predicted, independent of bias.

For verification as categorical variables, the continuous forecasts and observations, both rain rate and accumulation, were categorized into two categories, rain or no rain, and the forecast–observation pairs accumulated into 2 by 2 contingency tables according to the four possible outcomes: (a) hits (rain forecast and rain observed), (b) misses (rain observed but not forecast), (c) false alarms (rain forecast but not observed), and (d) correct negatives. To distinguish rain from no rain, a suitably low threshold was used, 1 mm h^{-1} for rain rate, and for accumulation, 0.005 mm per forecast period. To help evaluate the systems' performance on the heavier rain cases, the contingency tables were recomputed with higher thresholds, 5 and 20 mm h^{-1} for rain rate, and 1 and 5 mm per forecast period for the quantitative forecasts.

Various scores were computed from the elements of the contingency tables. The frequency bias is the ratio of the number of forecasts of rain to the number of rain observations:

$$\text{frequency bias} = \frac{\text{hits} + \text{false alarms}}{\text{hits} + \text{misses}}.$$

It measures an algorithm's tendency to over- or underforecast rain frequency or rain area. A value of 1 indicates an unbiased forecast, where rain is forecast as often as it is observed.

The critical success index (CSI), also known as the threat score, measures the ratio of correct rain forecasts to the total number of rain forecasts and observations:

$$\mathrm{CSI} = \frac{\text{hits}}{\text{hits} + \text{misses} + \text{false alarms}}.$$

A CSI of 1.0 indicates perfect correspondence between the forecast and observed rain occurrences. The CSI decreases when there is bias in the forecast rain area (increasing the frequency of misses or false alarms), and when there are errors in forecast rain location or timing (increasing both misses and false alarms).

A third score called the Hanssen–Kuipers (HK) discriminant, or true skill statistic, was computed for the verification against gauge data. This score is of the form

$$\mathrm{HK} = \frac{\text{hits}}{\text{hits} + \text{misses}} - \frac{\text{false alarms}}{\text{false alarms} + \text{correct negatives}}$$

and gives information about the utility of the forecasts for decision making. The first term, called the hit rate or the probability of detection, is the percentage of all rain cases that were correctly forecast. The second term, the false alarm rate, is the percentage of all no rain cases that were incorrectly forecast (false alarms). The HK is the difference of these two quantities, the range is +1 to −1, and a useless forecast scores a value of 0. The score measures the ability of the forecast system to separate the rain cases from the no-rain cases. Positive values are obtained if rain is forecast relatively more often when it occurs than when it does not.

Verification of the rain forecasts against both the radar analyses and the gauge observations gives a more comprehensive evaluation of the system performance characteristics. The analysis represent a uniform high-resolution data coverage in space, and the ‘‘observed’’ quantity, rain rate, is the same as the predicted quantity. Especially when compared against persistence, this evaluation gives a good estimate of the system’s ability to predict rainfall *patterns*. However, the analyses that are used as ‘‘observations’’ may themselves be biased, since they are subject to all the sources of error inherent in radar precipitation estimates. The gauge data, on the other hand, are more likely to provide completely *unbiased* verification data since they represent direct surface-based precipitation estimates. The gauge data are essentially point observations in space that are averaged in time, while the radar rain rates are instantaneous in time, but (nearly) continuous in space, at least to the pixel level. The matching strategy described above tries to preserve the integrity of both forecast and observation.

c. Boundary layer convergence lines

Both the location and length of nowcast convergence lines were verified. The forecast and analyzed convergence lines were represented in the FDP dataset as joined line segments. The location errors are determined by calculating, at each vertex in the forecast convergence line, the shortest distance to the adjacent analyzed convergence line. In cases where the forecast and analyzed lines had different lengths the ‘‘tails’’ (i.e., the segments present in one, but not both, line) are not used in the location error calculation. The shortest distances are then integrated over the length of the convergence line to obtain the mean and mean absolute location errors. The mean and mean absolute errors in predicted length are also calculated.

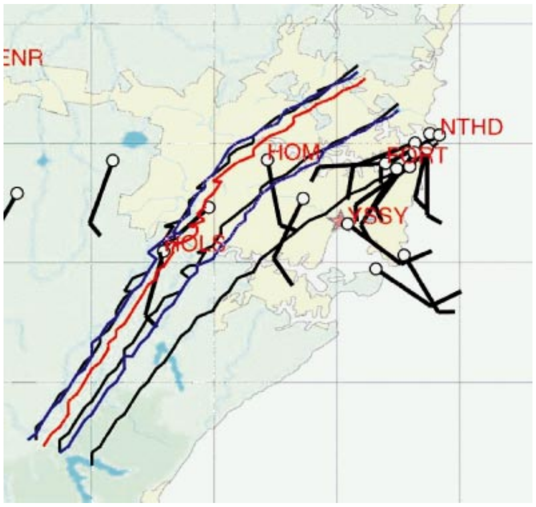

FIG. 3. The 30- (blue lines) and 60-min (red line) nowcasts for the sea-breeze front at 0130 UTC 20 Oct 2000, from the ANC algorithm. The black lines show the observed position.

d. Hail

The small sample available for comparison of severe weather leads us to concentrate on the hail reports for the 3 November case and make a qualitative assessment of performance of the two systems that assessed severe thunderstorm potential, namely CARDS and WDSS. Both algorithms’ forecasts are compared to the hail sizes from the observed reports. In addition, the temporal extent of large hail radar detections are compared to the time period of reports from the storm.

5. Verification of Auto-nowcaster

The Auto-nowcaster advects rain cells using either a reflectivity area tracker or steering-level winds derived from nearby soundings. Enhancement of convection is predicted to occur along boundary layer convergence lines and, particularly, where lines intersect. Although boundaries can now be detected automatically using variational analysis of the Doppler wind velocities and a mesoscale model (Sun and Crook 2001), in the FDP they were manually entered into the radar analysis. The boundaries are then extrapolated to produce 30- and 60-min nowcasts. Figure 3 shows an example of a nowcast for a sea-breeze front moving toward the northwest on 20 October 2000. The ANC made a very accurate prediction of its motion, with less than 2-km error in the 60-min forecast.

The mean and mean absolute errors in forecast boundary position are shown in Fig. 4 for 30- and 60-min forecasts. In a few isolated cases (not shown) the error

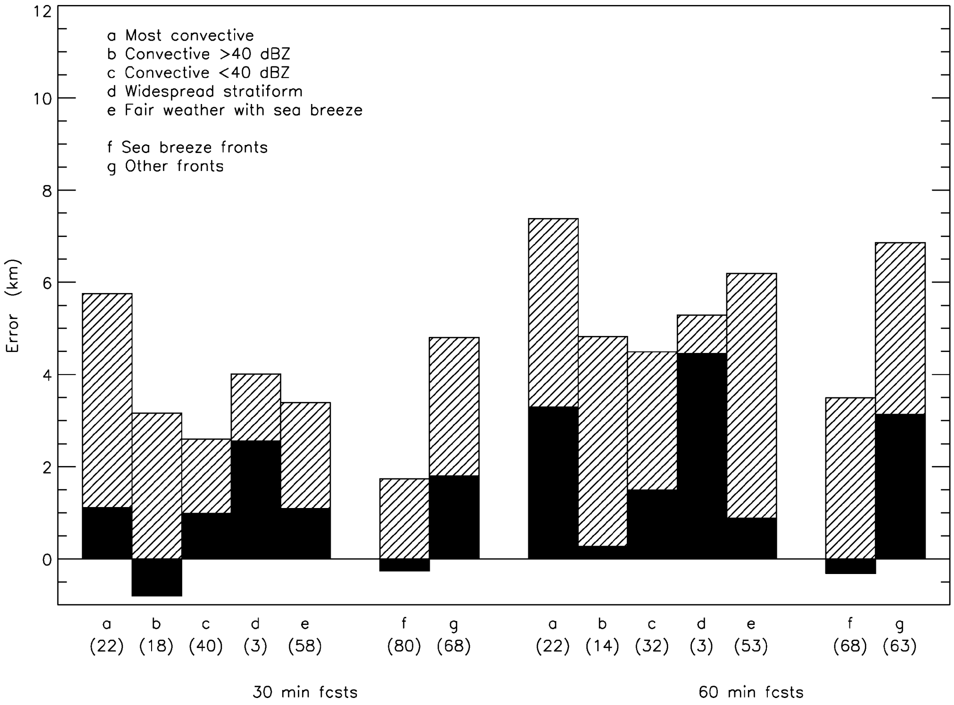

FIG. 4. Mean errors (solid) and mean absolute errors (hatched) in ANC forecasts of boundary position, as a function of weather type. The number in parentheses indicates the number of boundaries verified in each category.

in the forecast boundary position was as large as a few tens of kilometers after an hour, but the mean and mean absolute errors were much smaller. Looking first at the results for the weather types, the mean absolute errors were greatest for the most convective events where gust fronts moved rapidly through the domain. The errors were least for the days with light convection; the boundaries in these cases were sea-breeze fronts (see Table 4). Figure 4 shows that the MAE for sea-breeze fronts was half the value for other types of fronts. The mean error (bias) for sea-breeze front positions was very small, averaging −0.3 km. The forecast movement for other types of convergence lines averaged 3 km h^{-1} too fast.

The mean length of boundaries analyzed in the FDP was about 80 km. The length of the sea-breeze fronts did not change significantly over a 1-h period, but the average growth of gust fronts was about 15%. Figure 5 shows the mean absolute errors in forecast convergence line length, again stratified by weather type and front type. Discounting the stratiform category because of its small sample size, the mean absolute length errors averaged 6–18 km (7%–25%) after an hour. Not surprisingly, the length of sea-breeze fronts was better predicted than the length of gust fronts and other fronts.

The mean cell location error, determined using CRA verification, is plotted in Fig. 6 as a function of lead time. Errors were on the order of 10 km after 30 min. There was little overall difference in performance between the ANC and EXTRAP forecasts, and both performed much better than PERSIS.

Selected spatial verification statistics are shown as a function of lead time and rain threshold in Fig. 7. ANC had little bias for light and moderate rain, but showed a tendency to underforecast rain occurrence for rain exceeding 20 mm h^{-1}. The CSI for the nowcast and EXTRAP forecasts was about 0.3 after 30 min, decreasing to about 0.2 after an hour. The probability of detection for 30-min nowcasts in the FDP was 0.46, with a false alarm ratio of 0.51, an improvement over the values of 0.25 and 0.48, respectively, for 30-min nowcasts of Colorado storms (Wilson et al. 1998). The mean absolute errors were about 0.4 mm h^{-1}. The spatial correlation coefficient was 0.25 after 30 min and 0.15 after 60 min, slightly better than EXTRAP. ANC had lower MAEs than EXTRAP, but this is partly due to its tendency to underforecast heavy rain.

To test whether the differences in performance between ANC and EXTRAP were statistically significant, 95% confidence intervals were estimated using bootstrapping (Mason and Mimmack 1992). If the intervals overlap, then the two estimates cannot be considered significantly different. The confidence intervals for all measures were on the order of 1%–2% of the values themselves. The improvement of ANC over EXTRAP was statistically significant at the 95% confidence level only for the correlation coefficient and the mean absolute error (Fig. 7c).

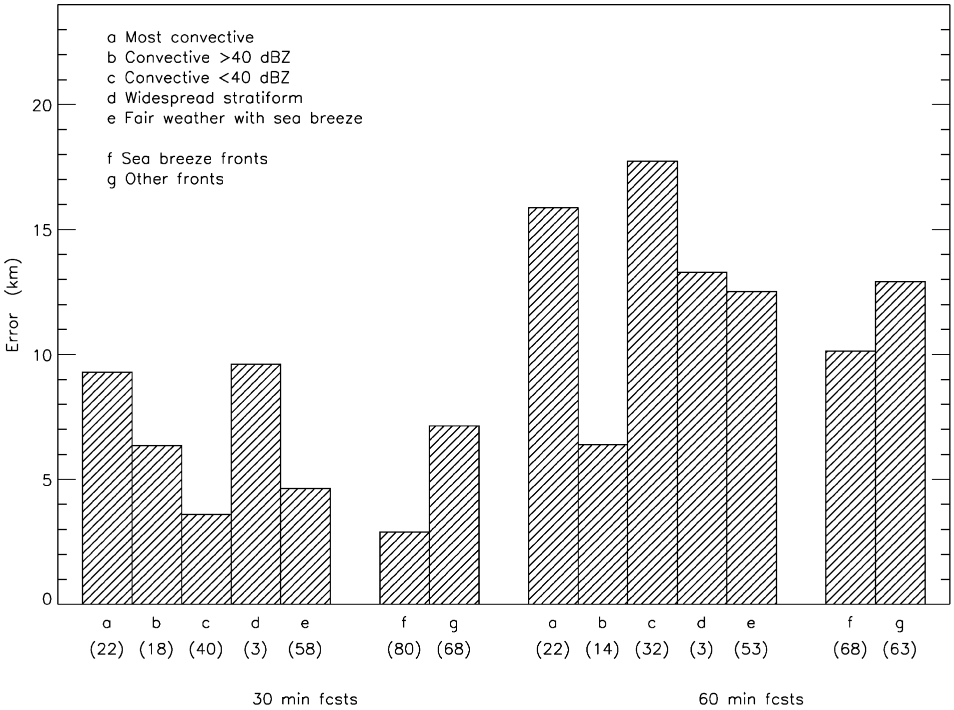

FIG. 5. Mean absolute errors in ANC forecasts of boundary length, as a function of weather type. The number in parentheses indicates the number of boundaries verified in each category.

The verification of ANC predictions of 30-min rainfall against gauge data is shown in Fig. 8. The scores at the initial time give a measure of how well the radar rainfall analyses agree with the gauge data. The apparent ‘‘errors’’ are due primarily to differences in spatial and temporal sampling, that is, the comparison of a snapshot for a 1-km radar pixel with an accumulation over 30 min at a point. These provide baseline scores against which forecast performance can be compared.

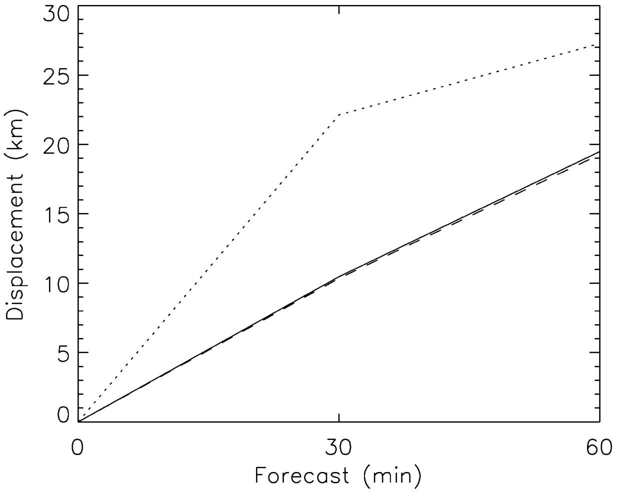

FIG. 6. Mean rain location error for the ANC algorithm (solid line), PERSIS (dotted line), and EXTRAP (dashed line), for all rain cases in Table 4.

Compared to rain gauge observations, the ANC underpredicted the frequency of rain, with the underprediction being more pronounced for higher rain amounts. Gauge-based PERSIS, with its natural scale advantage with respect to the verification data, outperformed ANC and EXTRAP both in terms of frequency bias and HK. The Hanssen and Kuipers score was slightly better for ANC than for EXTRAP for light rain rates (Fig. 8b). The mean bias was -0.15 mm for 60-min forecasts of half-hourly accumulation. The MAE-based skill with respect to persistence was quite low but improved with time as expected, going from -0.04 after 30 min to $+0.08$ after an hour. The extrapolation forecast was slightly more skillful in this respect.

6. Verification of CARDS

CARDS is a severe weather detection scheme that incorporated the TITAN cell display in the FDP. For the rain forecast CARDS used the original nowcasting technique of Bellon and Austin (1978) where the domain is subdivided into 3 × 3 subdomains and the latest constant altitude plan position indicator (CAPPI) image is extrapolated by the nine velocity vectors. Rain rates were predicted at 11 sites in the Sydney region.

The CARDS system produced two estimates of hail size, an average size and a maximum size for each radar volume scan every 5 min. Detections associated with the 3 November 2002 storm began on 0330 UTC and continued until 0605 UTC (Fig. 9). Estimated hail size

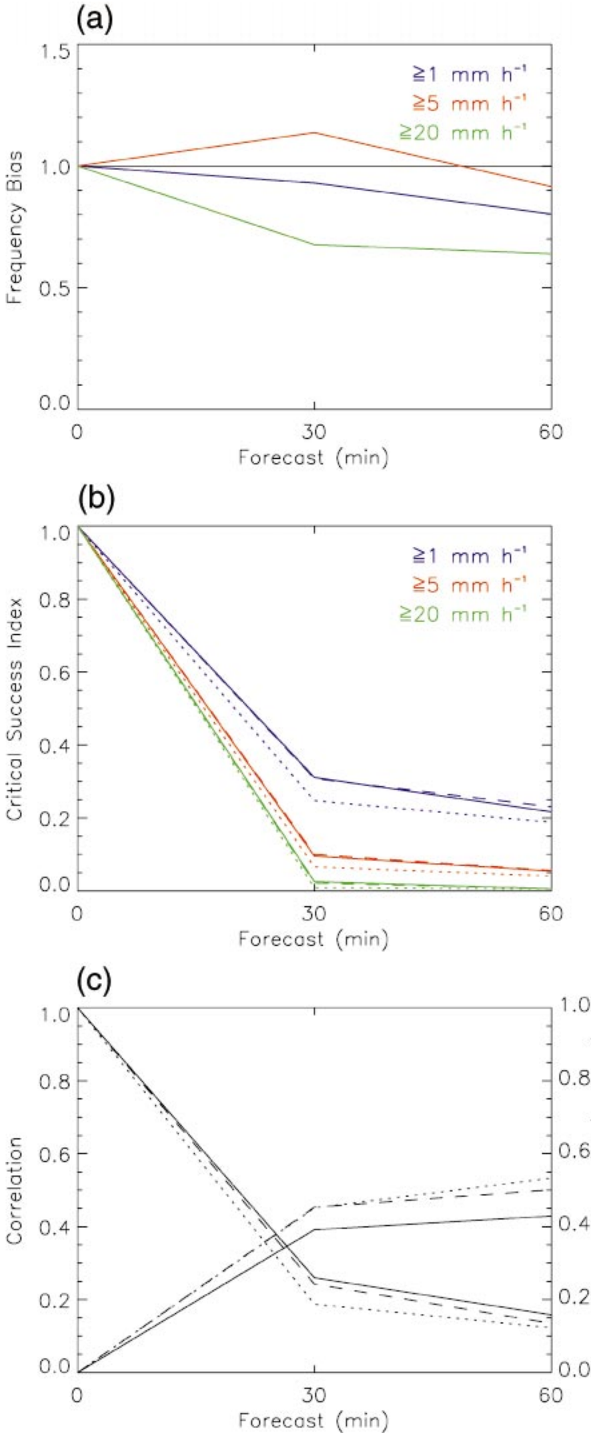

FIG. 7. (a) Frequency bias, (b) CSI, and (c) MAE and spatial correlation coefficient for the ANC algorithm (solid line), PERSIS (dotted line), and EXTRAP (dashed line), as verified against radar analyses, for all rain cases in Table 4.

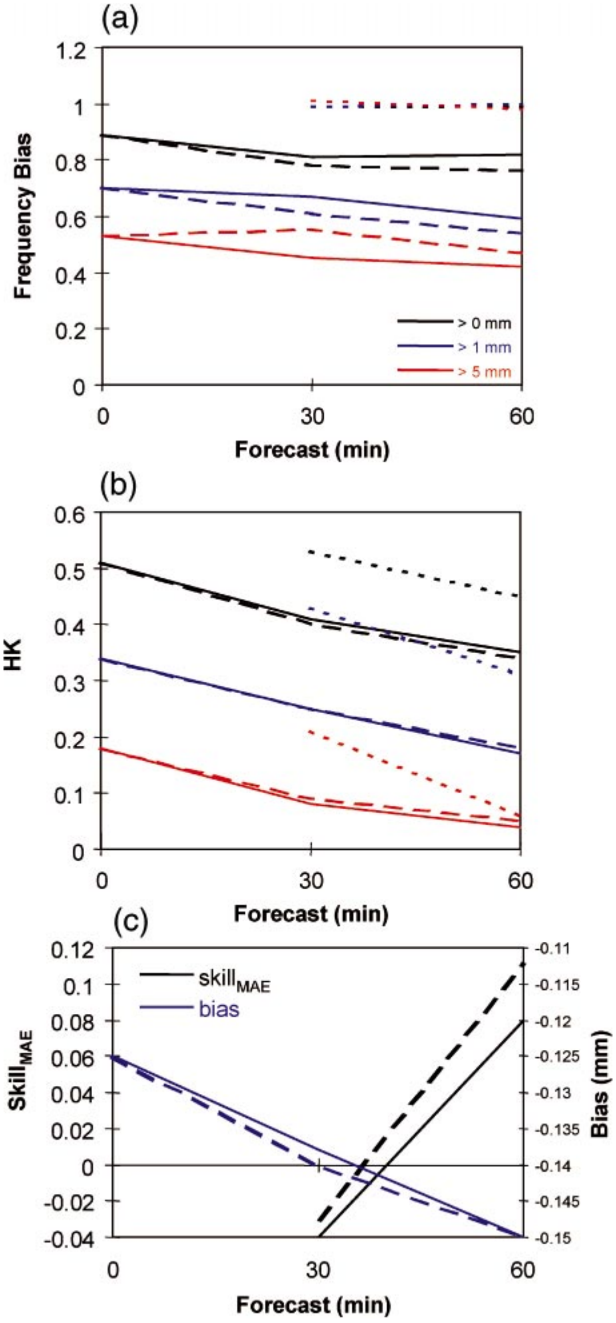

FIG. 8. (a) Frequency bias, (b) HK score, and (c) linear bias and MAE-based skill with respect to persistence for the ANC algorithm (solid line), PERSIS (dotted line), and EXTRAP (dashed line), as verified against rain gauge data, for all rain cases in Table 4. In (a) and (b) the black line corresponds to all rain, the blue to rain ≥1 mm, and the red to rain ≥5 mm.

grew rapidly from less than 1 cm to 6 cm or more for the maximum hail size in approximately half an hour and to about 2 cm for the average hail size over that time period. High values were observed consistently from 0400 to 0600 UTC, after which time the detections ended abruptly. In order to smooth the radar estimates in time, a three-point boxcar median filter was applied, followed by a five-point running mean, resulting in a relatively smooth signal. The smoothed estimates suggest that there were two maxima detected, one shortly

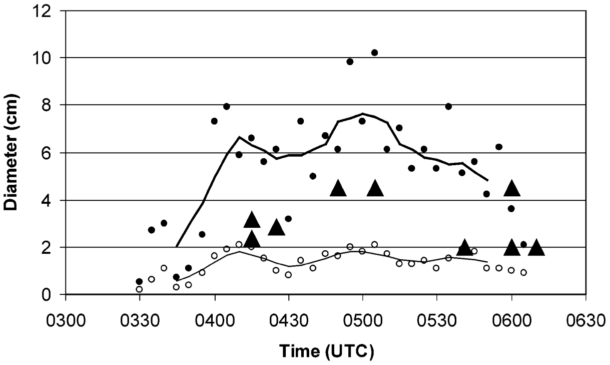

FIG. 9. Time series of estimated hail size from CARDS for the 3 Nov 2000 storm. Filled (open) circles are raw maximum (average) hail size in cm. Thick (thin) line is smoothed maximum (average) hail size. Triangles show reported hail size.

after 0400 UTC and a second between about 0445 and 0510 UTC.

When the radar estimates are compared to the reported hail size, we see that the reported sizes generally fall between the maximum and average estimated sizes. There is a lag at the beginning of the period after first detection by radar before the first report. This could be an artifact of reporting problems, but it is also possible that it is physically real. Radar detections necessarily occur first above the ground. Hailstones take time to fall, so that we would expect, even with perfect reporting, to see the reports lag the detections.

The other item of interest concerns the Greystanes report of 4.5-cm hail at 0600 UTC. This report seems to be late compared to the track of the storm. If it is moved 30 min earlier in time, consistent with the track, it would provide a third large hail report in a series from 0450 to 0530 UTC. If so, then the impression given by the reports is of a period of relatively small hail (2–3-cm diameter), followed by much larger hail, and, finally, at the end of the storm, relatively small hail (2 cm). The qualitative impression is that the CARDS system provides an envelope in time, given the lag between hail formation aloft and hail reaching the ground, and size of hail for this event. Lead times from first detection of 2-cm hail to first report, and from first detection of 4-cm hail to first report, are both on the order of half an hour. If this performance could be sustained over a larger dataset, the hail detection would certainly be of value for forecasters.

CARDS made rainfall predictions for 11 sites, but unfortunately verifying gauge data were easily available for only two of them, Fairfield and Ryde. However, the long time series of observations led to a large variety of weather being sampled, so we expect that the results for the two sites will not be unrepresentative of the overall performance of CARDS for the Sydney domain. EXTRAP forecasts were not available for this algorithm.

CARDS overpredicted the frequency of light rain, but underpredicted the frequency of heavier rain (Fig. 10a), with this tendency increasing with time. Encouragingly, its skill at predicting rain occurrence, as measured by the HK score, was greater than PERSIS for both 30- and 60-min forecasts of heavier rain, although this was not the case for lighter rainfall. The decline in performance with increasing rain rate was also found by Bellon and Austin (1978), who reported CSI values of 0.33 and 0.12 for rain exceeding 1 mm h^{-1} and 5 mm h^{-1}, respectively, for storms near Montreal during the sum-

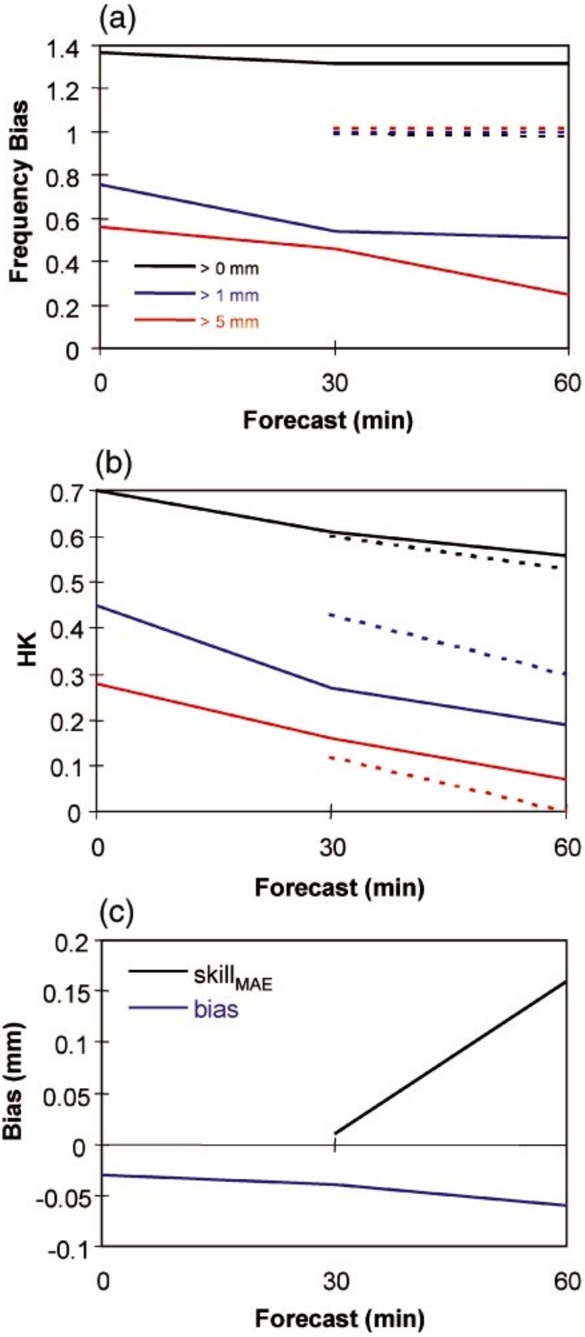

FIG. 10. As in Fig. 8 but for the CARDS algorithm.

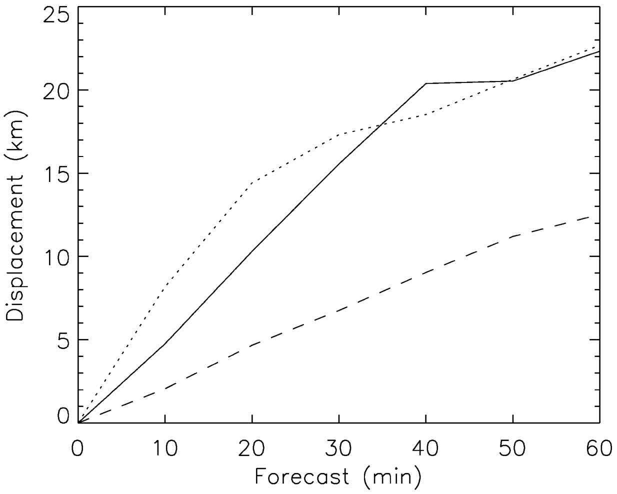

FIG. 11. As in Fig. 6 but for the Gandolf algorithm.

mer of 1976. Those nowcasts were made at a spatial resolution of about 6 km, and verification was performed against radar analyses as opposed to gauge observations. The Skill$_{MAE}$ was positive for both forecast periods, and CARDS had a small negative mean bias of -0.05 mm for 30-min accumulations.

7. Verification of Gandolf

The Gandolf algorithm diagnoses and predicts the life cycle of individual thunderstorms, based on a combination of radar, satellite, and numerical weather prediction (NWP) data. Storms are represented in the radar analyses and nowcasts as square cells with enhanced rainfall. The cells are horizontally advected according to a representative NWP model wind vector. The version of the Gandolf algorithm used in Sydney ran at a spatial resolution of 2 km.

The mean location error for Gandolf rain nowcasts is shown in Fig. 11. As discussed by Pierce et al. (2004, in this issue) the scheme had an unfortunate tendency to rapidly decay the rain cells, resulting in few cells remaining after 30 min, and even fewer after 60 min. While the performance of Gandolf was better than PERSIS early in the forecast period, it was poorer than EXTRAP for the same set of cases.

The bias score in Fig. 12a clearly illustrates Gandolf's tendency to shrink rain cells. The predicted frequency of moderate and heavy rain was much smaller than observed after the first 10 min, and virtually nonexistent after 30 min. This leads us to suspect that the nowcast scheme had not been properly tuned for the NWP model input and/or the Sydney conditions, since Gandolf nowcasts have previously performed well in the United Kingdom. According to Pierce et al. (2000) Gandolf achieved CSI values of 0.47 for 30-min nowcasts of 15-min accumulated rainfall in a 2.5 km^2 catchment during the summers of 1995 and 1996. During the FDP it dissipated the rain cells when the atmospheric sounding was stable, which was for most of the experiment. Gandolf's decaying tendency was also responsible for its poor performance in predicting the occurrence of rain of any magnitude beyond the first 10 or 20 min (Fig. 12b), and its poor spatial correlation coefficients for 30

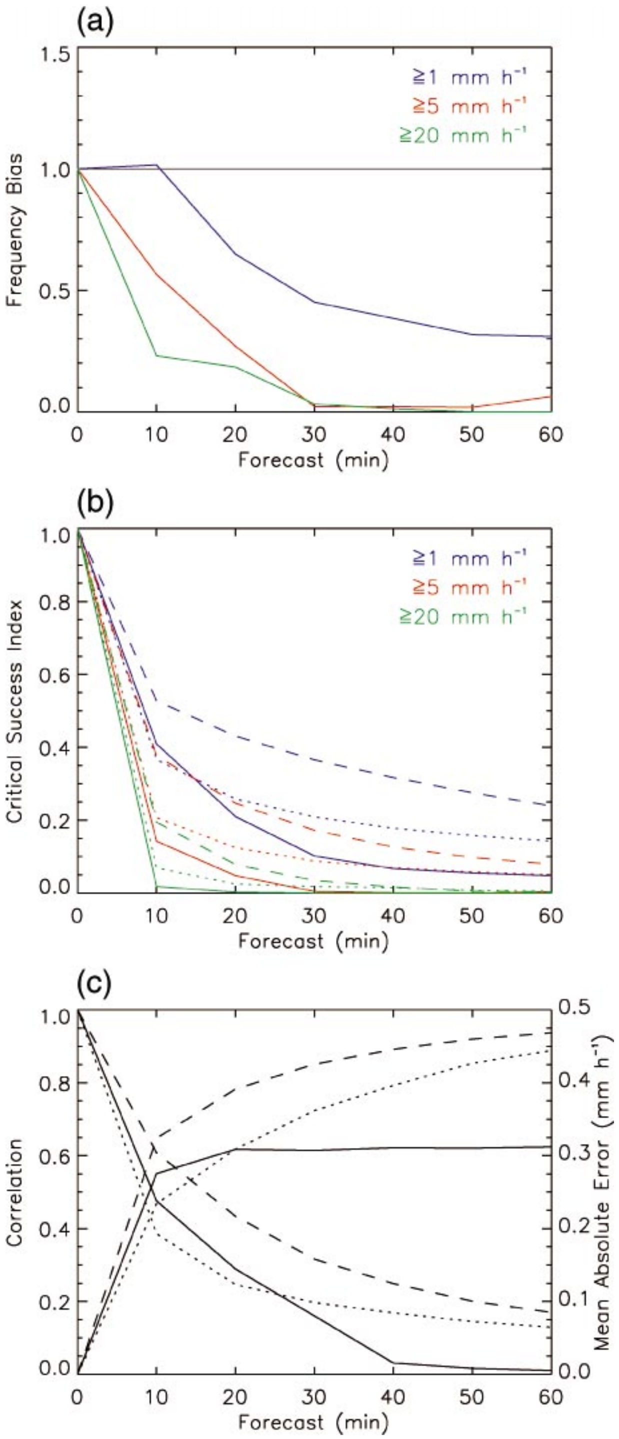

FIG. 12. As in Fig. 7 but for the Gandolf algorithm.

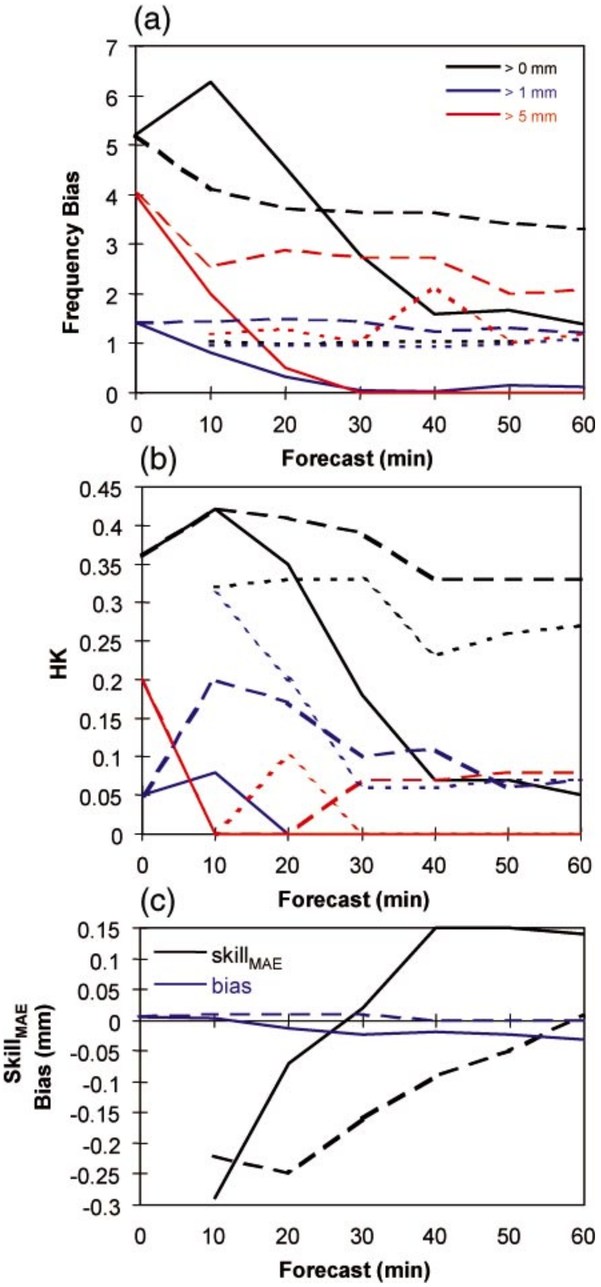

FIG. 13. As in Fig. 8 but for the Gandolf algorithm.

min and beyond (Fig. 12c). The mean absolute error remained about 0.3 mm h^{-1} for the whole forecast period.

The gauge verification in Fig. 13 is for 10-min rain accumulations. The noisiness of the figure reflects the fact that many Gandolf forecasts were missing, especially during the most convective events. The Gandolf analyses had a much greater frequency of rain >0 mm and rain >5 mm than was observed at the gauge sites. The frequency of heavy rain rates decreased quickly during the forecast, but the frequency of very small amounts was overforecast throughout the period. The Hanssen and Kuipers scores (Fig. 13b) show that after the first 10 min Gandolf did not perform as well as PERSIS or EXTRAP at predicting rain occurrence. However, in an MAE sense (Fig. 13c), Gandolf did show greater skill than PERSIS later in the forecast because it did not make erroneous predictions of heavy rain. The mean error of the algorithm was -0.03 mm 10 min^{-1} for a 1-h forecast.

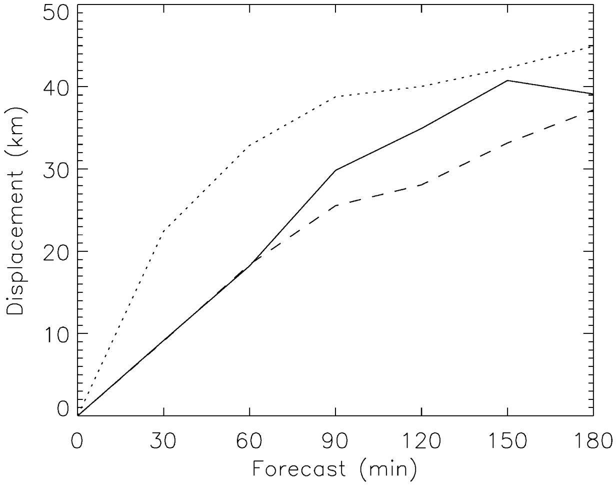

FIG. 14. As in Fig. 6 but for the Nimrod algorithm.

8. Verification of Nimrod

The Nimrod algorithm is unique among the nowcast algorithms used in the FDP, in that it extends the forecast period beyond the usual 60 min by blending the nowcast into the forecast from a mesoscale NWP model. This means that nowcasts at 30 and 60 min, while representing mainly the advection of radar-detected rain cells, have a small component of NWP model forecast in them. This scheme was run at a fairly coarse spatial resolution of 5 km.

The mean location error is plotted as a function of time in Fig. 14. The location errors for Nimrod forecasts averaged 9 km after 30 min, roughly the same as EXTRAP and much better than PERSIS. However, at 90 min and beyond, Nimrod forecasts had greater location errors than EXTRAP, possibly reflecting the influence of errors in the NWP model forecasts.

Nimrod showed a slight positive bias in predicting the frequency of light and moderate rain, as shown in Fig. 15a, but the frequency of heavy rain was underpredicted by 20%–30% throughout the forecast period. The CSI (Fig. 15b) showed that Nimrod had some skill in predicting light rain occurrence out to at least 3 h, while the skill for rain exceeding 20 mm h^{-1} was exhausted in the first hour. The performance of Nimrod was comparable to that of EXTRAP during the first 60 min, which is not surprising since the Nimrod forecasts are heavily weighted toward radar-based advection. The

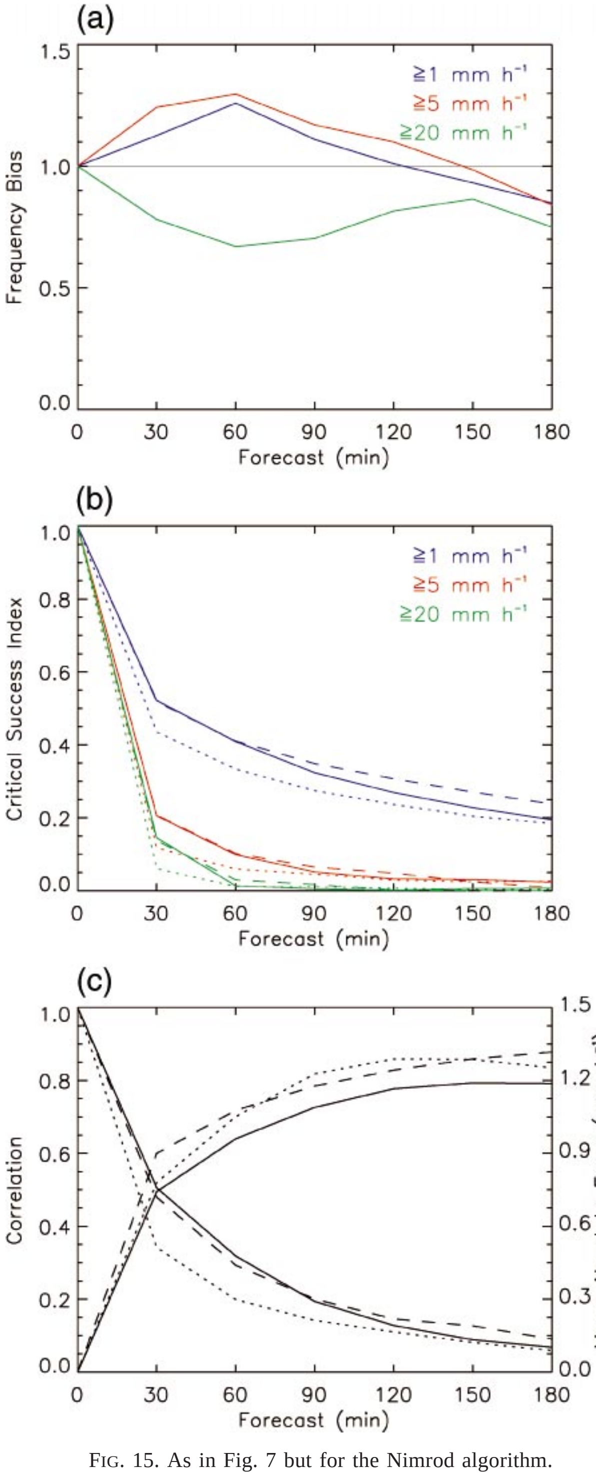

FIG. 15. As in Fig. 7 but for the Nimrod algorithm.

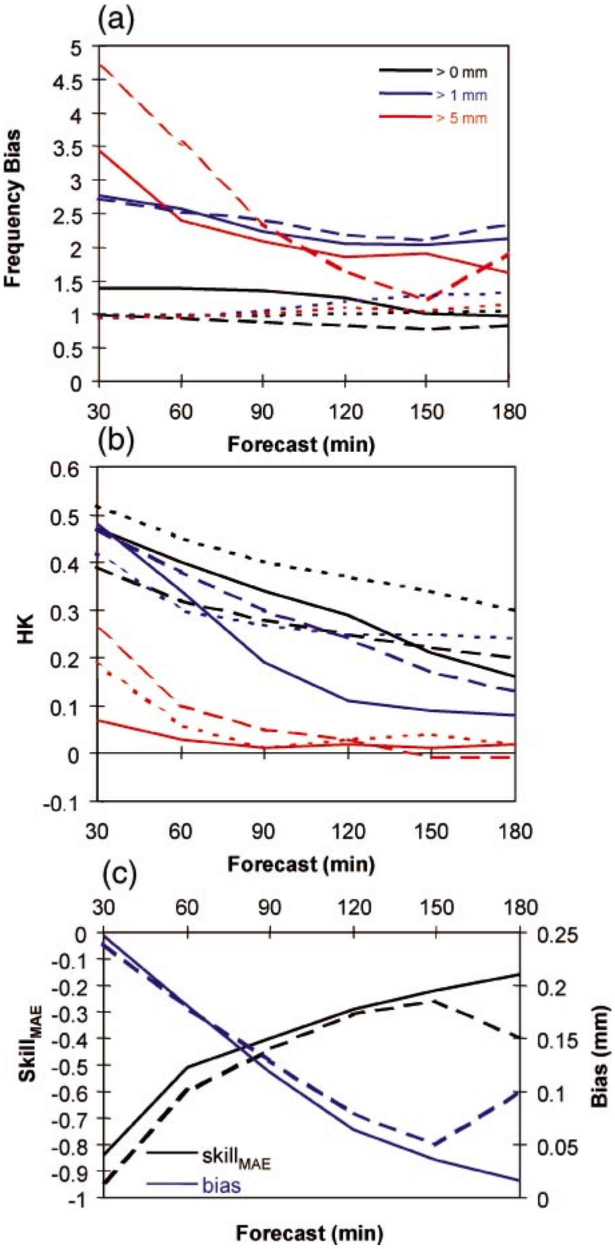

FIG. 16. As in Fig. 8 but for the Nimrod algorithm.

probability of detection for 60-min Nimrod nowcasts during the FDP was 0.66 and the false alarm ratio was 0.48, slightly poorer than the values of 0.77 and 0.26, respectively, reported for a 40-day test period in the United Kingdom during the summer of 1995 (Golding 1998). Nimrod had mean absolute errors that were fairly high, about 1 mm h^{-1} for 60-min forecasts. Its MAE and spatial correlation coefficients appeared to be slightly better than EXTRAP. However, none of the differences between Nimrod and EXTRAP shown in Fig. 15 were significant at the 95% level.

When compared to the gauge observations, Nimrod had very large frequency biases throughout the forecast period, particularly for the higher rain rates (Fig. 16a). This may be partly due to its coarse spatial resolution, since the occurrence of rain anywhere within the grid box would be represented by a 25 km^2 region of rain. It is also related to a broader overestimation of spatial

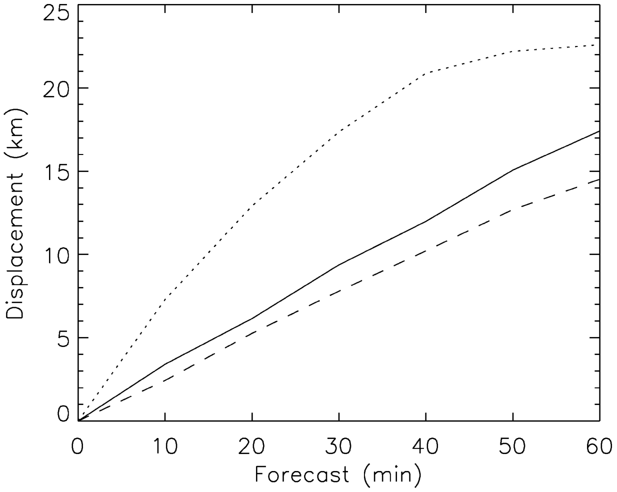

FIG. 17. As in Fig. 6 but for the S-PROG algorithm.

rain rates (e.g., see Fig. 1). Figure 16c shows that the mean bias for 30-min nowcasts of half-hourly rain is 0.25 mm. The bias decreases with time as an increasingly greater fraction of the forecast is derived from the NWP model, so that after 3 h the bias is close to 0. The large bias is also responsible for the MAE-based skill score being negative for at least 3 h into the forecast. The occurrence of rain >0 mm was better predicted by PERSIS than by Nimrod or EXTRAP (Fig. 16b), a result of the high false alarm rate in the radar-based forecasts. EXTRAP provided the best forecasts of heavier rainfall occurrence for the first 2 h.

9. Verification of S-PROG

S-PROG, short for Spectral Prognosis, is an advection based algorithm that represents rain as a multiplicative cascade of structures characterizing a spectrum of spatial and temporal scales (Seed and Keenan 2001). The version used in the Sydney FDP assumes that the total rain area remains constant in time, with the smaller-scale structures (usually corresponding to the more intense rain cells) decaying more quickly than the larger-scale structures. This produces an increasingly smooth field as the forecast progresses.[2]

Figure 17 shows the mean rain location errors for the S-PROG algorithm. It had slightly greater location errors than EXTRAP nowcasts throughout the forecast period, with values of about 10 km after 30 min. The difference was most likely due to the influence of S-PROG's smoothing of rain maxima on the CRA pattern matching results.

The smoothing is also evident in the plot of bias score versus time (Fig. 18a). Few occurrences of rain greater than 20 mm h^{-1} remained after only 10 min, and after 60 min the frequency of rain exceeding 5 mm h^{-1} was also severely underestimated. The CSI plot (Fig. 18b) shows that for the duration of the forecast period S-PROG provided slightly better predictions of light rain occurrence than did EXTRAP and PERSIS, with a value of 0.25 after 60 min. Performance for moderate rain was

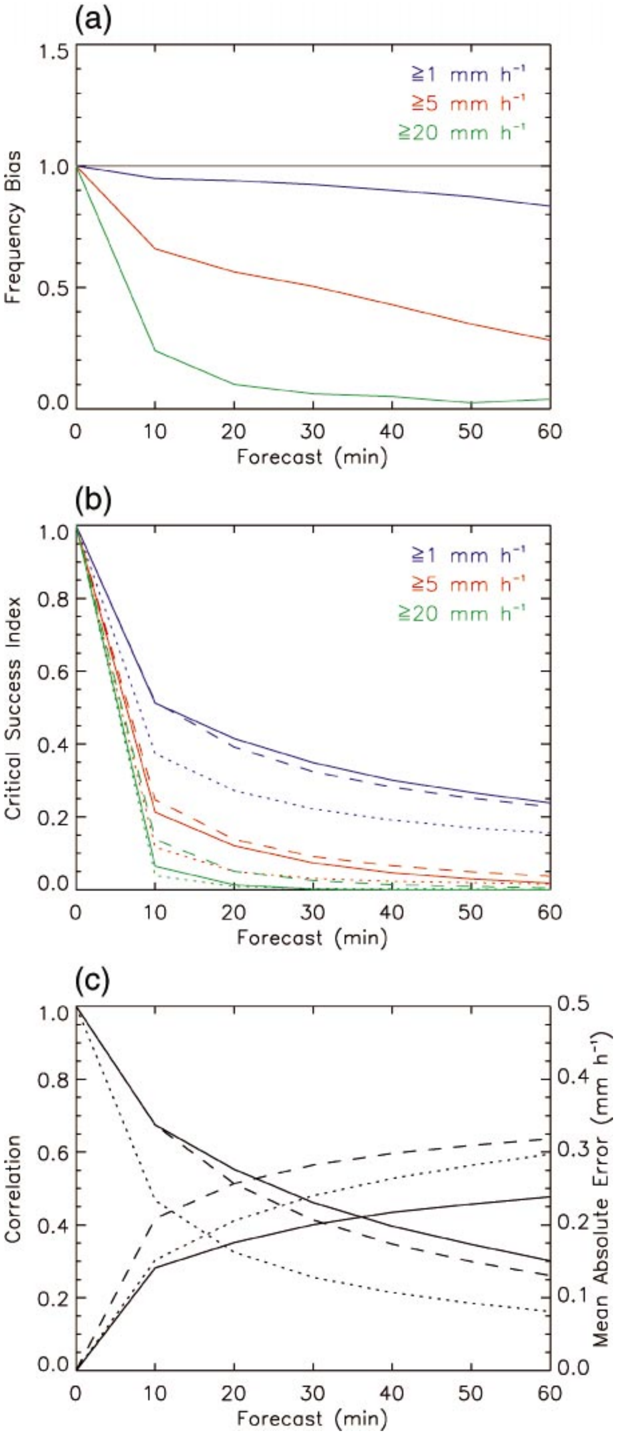

FIG. 18. As in Fig. 7 but for the S-PROG algorithm.

[2] A more recent version of S-PROG also requires the mean rain rate to remain constant with time.

slightly worse than that of EXTRAP, while performance for heavy rain was significantly worse. The occurrence of convective rain was predicted with greater skill than stratiform rain. In terms of both the mean absolute error and the spatial correlation coefficient, S-PROG performed notably better than EXTRAP and PERSIS (Fig. 18c). This is the result of the conservative nature of S-PROG forecasts, and is consistent with the scheme's philosophy of not attempting to predict the smaller-scale structures beyond their natural timescales. Because of the large number of samples, the 95% confidence intervals for the verification statistics were very tight, and all differences between S-PROG and EXTRAP were statistically significant at the 95% level.

The verification of S-PROG rain nowcasts against the observed 10-min accumulation at rain gauges is presented in Fig. 19. The radar analysis produced excessive very light rain and insufficient heavier rain when compared to the gauge data, and this was carried through to the forecasts. As was seen in Fig. 18, the algorithm decayed the heavier rain rates quickly with time. S-PROG outperformed both EXTRAP and gauge-based PERSIS in predicting the occurrence of rain >0 mm (Fig. 19b). The opposite was true for the heavier rain rates due to the low hit rate. The MAE-based skill with respect to persistence was positive for both S-PROG and EXTRAP, reaching 0.25 after an hour. The mean bias was negative throughout the forecast period at about -0.06 mm 10 min^{-1}. Even though the heavy rain rates decayed quickly, the mean bias decreased only slowly because most of the rain was of low intensity.

10. Verification of TITAN

TITAN is a cell tracking algorithm in which the intensity, speed, and direction of an individual thunderstorm are determined from its past history. In the FDP each TITAN rain cell was represented as an idealized ellipse whose area was equal to the area enclosed by the 35-dB*Z* isoline, and whose rain rate was constant at 5.6 mm h^{-1}. TITAN can also forecast cell growth or decay, but this ''trending'' feature was not applied in the FDP. Rather, to suggest uncertainty in cell movement, the major and minor axes of each ellipse were systematically increased by 3 km h^{-1}. The verification results pertain to this highly simplified ''cartoon'' version of TITAN (see Fig. 1). TITAN was run using reflectivities from two different radars, the C-band Doppler radar near Sydney Airport, which had 5-min sampling, and an S-band weather radar in Wollongong, with 10-min sampling. The results shown in this section are for the TITAN forecasts made using the Sydney radar data.

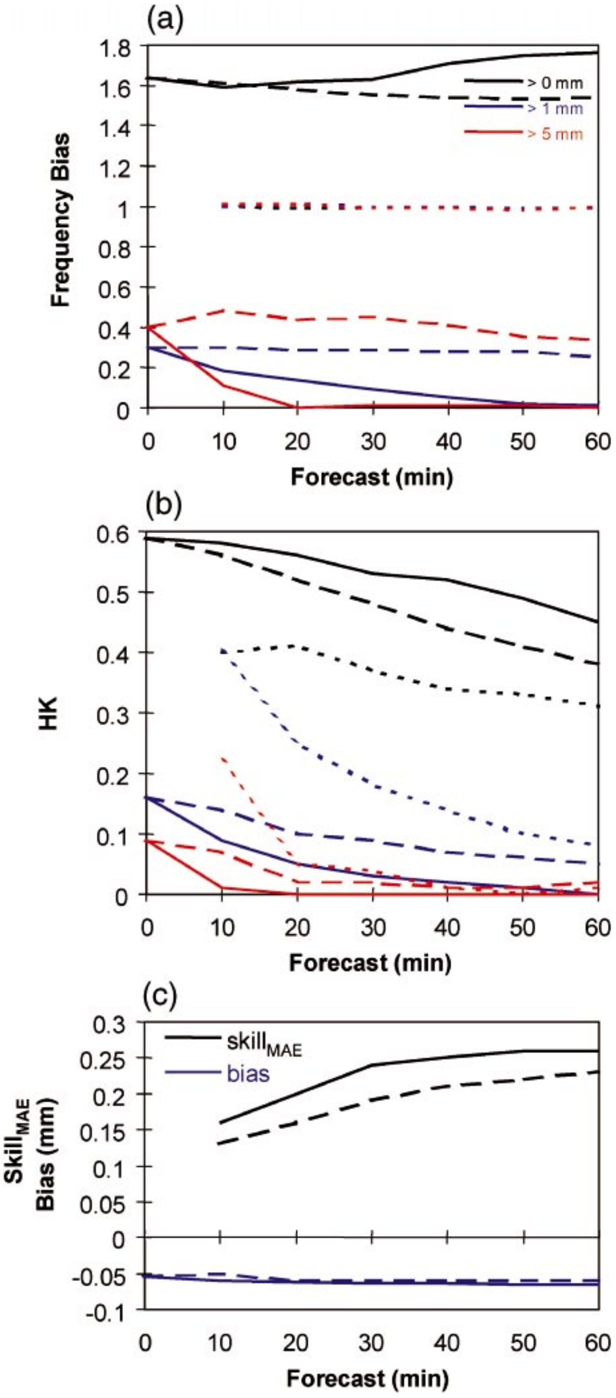

FIG. 19. As in Fig. 8 but for the S-PROG algorithm.

The cell position errors for TITAN are compared to EXTRAP for two ranges of cell intensity in Fig. 20. Although the typical position errors were similar for both forecasts, the mean and median errors were consistently smaller for TITAN than for EXTRAP. This demonstrates the advantage that cell trackers have over area trackers in allowing the motion of individual storms to deviate from the mean rain field motion. The mean position errors were slightly lower for the more intense cells than for the less intense cells, with fewer large errors. This is not surprising since the storms with greater reflectivity are generally larger, longer lived, and

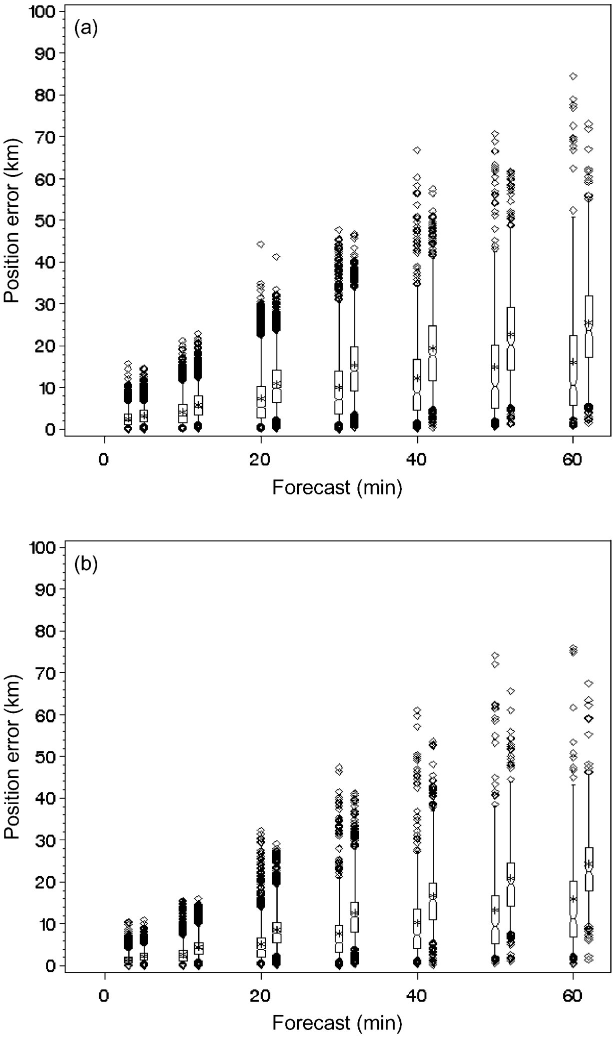

FIG. 20. Box plots of cell position errors for TITAN nowcasts (left member in each pair) and EXTRAP (right member), for two ranges of maximum cell intensity: (a) 30–50 and (b) >50 dB*Z*. The asterisk indicates the mean value, the boxes indicate the 25th, median, and 75th percentiles, and the lines indicate the 10th and 90th percentiles. Although forecasts were made every 5 min, values are shown only every 10 min in the figure.

therefore easier to track. The median (mean) track error after an hour was 10 km (15 km). The TITAN forecasts from the Wollongong radar with 10-min sampling had a mean error of 18 km, illustrating the benefit of more frequent sampling.

The spatial verification for the idealized TITAN nowcasts should be considered as cell, as opposed to rain rate, verification. The imposed cell growth is reflected in the linear increase in bias score with time, as shown in Fig. 21a. The values of CSI for TITAN and EXTRAP were 0.25 after 30 min and 0.15 after 60 min, significantly better than PERSIS. These values correspond quite closely to the CSI value of 0.22 found by Brown and Brandes (1997) for 30-min 2-km resolution TITAN nowcasts for 11 convective storm days in 1994 and 1996, and a CSI of 0.25 for 30-min 5-km resolution nowcasts in Colorado during the summers of 1989 and 1990 (Dixon and Wiener 1993).

The TITAN algorithm in this cartoon form should not be used to predict quantitative rainfall, so no verification against the rain gauge data was done.

11. Verification of WDSS

WDSS generates probability of severe hail (diameter ≥2 cm), probability of hail, and a maximum expected hail size. The time series of the probability of the hail product indicates the WDSS first gave a nonzero probability of hail with the 3 November 2000 storm at 0350 UTC (Fig. 22). By 0400 UTC, the hail probability was 100% and remained at that level for all but two volume scans (0420 and 0430 UTC) through 0555 UTC after which it rapidly declined to 0% by 0610 UTC. The first nonzero probability of severe hail (hail of at least 2-cm diameter) (POSH) was at 0400 UTC. The POSH remained at 80% or higher for all but one volume scan through 0555 UTC. Using a three-point median, five-point running mean smoother on POSH, the maximum around 0500 UTC is apparent. Taking into account the lag between hail formation and hail on the ground, high values of POSH describe the period of time during which the largest hail fell very well. When the time discrepancy of the Greystanes report is accounted for, the time series of POSH and report size match up well. This is in agreement with the more extensive study of Witt et al. (1998), where the probability of detection for severe hail from the WDSS hail algorithm was about 70%.

WDSS also produces a maximum expected hail size.[3] Although the time series of the maximum size shows a maximum between 0500 and 0530 UTC that might correspond to the times of the largest reports (Fig. 23), it also shows a maximum at the earliest stages that values are estimated, around 0410 UTC. In general, after correcting the time of the Greystanes report, the decrease in hail size at the end of the storm is captured well, but the apparent maximum in the middle of the storm lifetime is not as obvious in the estimated size as it is in the reports. It is important to remember the caveats about the accuracy of reports, though. Much larger sample sizes would be needed to evaluate the forecast hail size adequately.

The WDSS Storm Cell Identification and Tracking (SCIT) algorithm tracks cell cores, defined by the innermost 3D contour of reflectivity in consecutive volume scans [contour interval of 5 dB*Z*; Johnson et al. (1998)]. The cell position errors are plotted as a function of lead time in Fig. 24. The median track error for the

[3] The values shown here are the correct values produced by the algorithm. Values that were available to the Sydney 2000 forecasters in real time differed from these and were in error due to a coding error in transferring information from the algorithm to the text display.

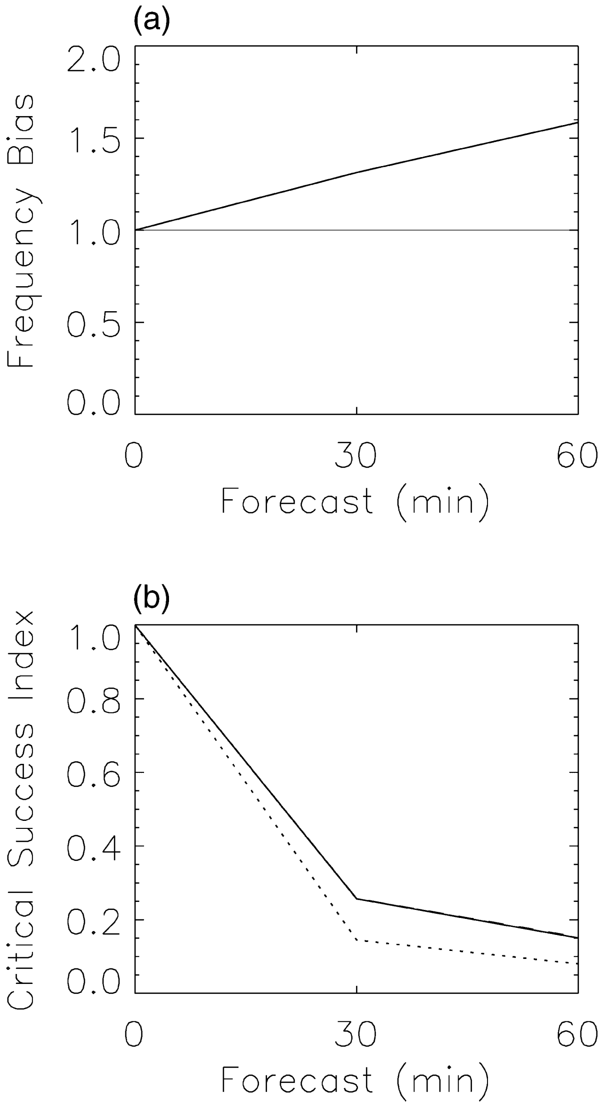

FIG. 21. (a) Frequency bias and (b) CSI for the idealized TITAN algorithm (solid line), PERSIS (dotted line), and EXTRAP (dashed line), as verified against radar analyses, for all rain cases in Table 4. Note in (a) that the values of PERSIS and EXTRAP are 1.0 and in (b) the dashed EXTRAP line overlays the solid TITAN line.

most intense cells is 7 km after 30 min and 14 km after an hour. The mean values are much higher (14 and 34 km, respectively), the result of some very large individual errors that were most likely caused by incorrect association of cell cores in consecutive radar analyses. For weaker cells the mean track errors are greater because they typically have shorter life spans, and because their motion is more difficult to diagnose. By way of comparison, Johnson et al. (1998) reported mean track errors of about 10 km after 30 min and about 23 km after 60 min for 17 storms in the United States during 1992–95.

Compared with EXTRAP, the WDSS nowcasts have lower track errors for the most intense cells but greater errors for the weak cells. This supports many earlier findings that cell trackers are best suited for the most convective situations because individual thunderstorm motion can often deviate from the mean flow.

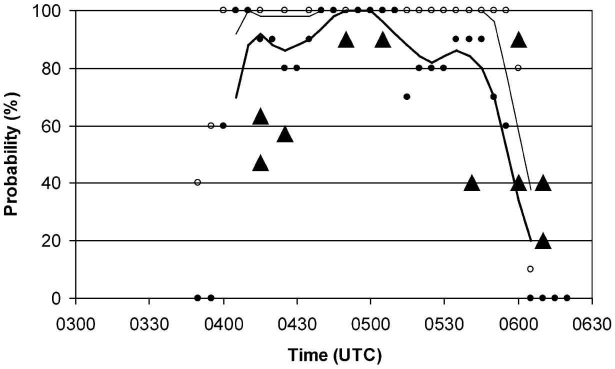

FIG. 22. Time series of the probability of hail from WDSS for the 3 Nov 2000 storm. Filled (open) circles are raw probability of severe (any) hail size in %. Thick (thin) line is smoothed probability of severe (any) hail. Triangles show reported hail size in cm multiplied by 20.

12. Verification of TIFS

The Thunderstorm Interactive Forecast System (TIFS) is an interactive tool that allows the user to modify the details of any storm displayed by a cell-based nowcast algorithm. The location, speed, direction, shape, size, and intensity of a cell are all attributes that can be easily changed, and the graphical output can be annotated to make it simple for external users to understand [Bally (2004) describes the TIFS system in detail]. TIFS used TITAN nowcasts as the starting point in the majority of situations during the FDP. Forecasters made modifications to the nowcasts, usually to remove uninteresting or erroneous cells, occasionally to change the speed and direction of the cells of interest, and to draw storm warning areas. The modified graphical output was then sent electronically to external users. Com-

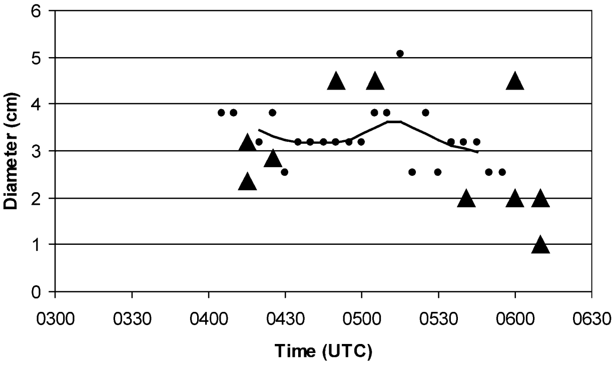

FIG. 23. Time series of maximum expected hail size from WDSS for the 3 Nov 2000 storm. The filled circles are raw maximum hail size in cm and the line is the smoothed maximum hail size. Triangles show reported hail size.

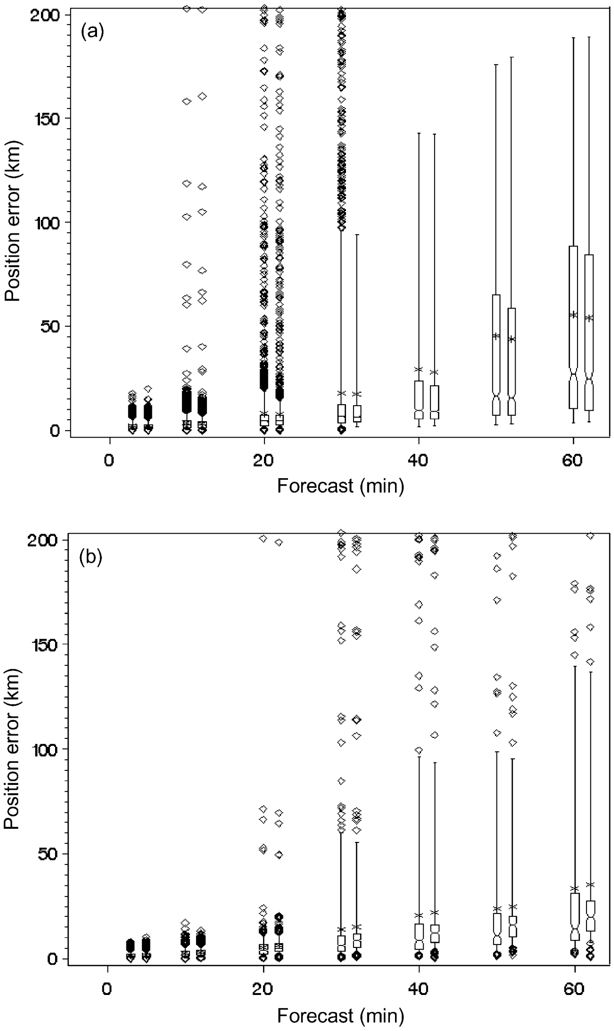

FIG. 24. As in Fig. 20 but for the WDSS algorithm.

parison of the verification results for the modified and unmodified forecasts can help to answer question 6 in the introduction, ‘‘Do the forecasters improve the quality of the forecasts compared to the ‘raw’ FDP products alone?’’

Twenty-three TIFS nowcasts were issued on 5 days during the FDP, marked by asterisks in Table 4. They all corresponded to periods of significant convection that would be of concern to the external users. The forecasters used TIFS to manually filter out about two-thirds of the cells that they considered less important or incorrect. The speed and direction were modified in only 4% of the remaining cell tracks, suggesting that the forecasters were usually satisfied with the quality of the automated tracks for those cells.

The location errors for TIFS and the original (unfiltered) nowcasts are plotted as a function of lead time in Fig. 25. Only those cells that were visible in the products sent to users are included in the TIFS verification. As seen previously for the cell tracking algorithms, the track errors increase with time and are greater for the shallow cells than for the more intense cells. The greatest improvement made by the forecasters was the removal of unimportant and obviously erroneous cells, which significantly reduced the mean track error for the moderate and intense cells. Examination of the few tracks that were modified indicated that their mean errors were 20% greater than those that were unmodified, suggesting that the automated track forecasts for these generally more intense and longer-lived cells might better be left alone.

13. Conclusions

The quantitative verification results shown in this paper provide some answers to the questions posed at the start of the Forecast Demonstration Project. In interpreting these results it is important to bear in mind that (a) the results apply to springtime weather in the region surrounding Sydney, Australia, and that extrapolation to other situations must be done with caution; (b) in general the algorithms were not optimally tuned for the Sydney conditions, and some were missing certain types of input data; (c) the verification data were imperfect in terms of spatial and temporal resolution, bias, radar-related quality control issues, and extent and frequency of hail reports; and (d) only the most ‘‘interesting’’ subset of the full FDP dataset was verified.

1) Is it feasible to predict the location of convection with enough accuracy and skill to be useful?

   The nowcast algorithms were able to successfully predict the location of the most convective cells with about 15–30-km mean error, and 10–14-km median error, for 1-h forecasts. Since the speed of these systems was typically 60 km h^{-1}, the median errors are on the order of 15%–25% of the distance traveled. The track errors for less intense cells were slightly greater than for the most intense cells, due to their less organized nature. The positive values of the critical success index and spatial correlation coefficient for 60-min forecasts confirm that all of the algorithms had skill in predicting the location of the cells. However, only the cell tracking algorithms were able to outperform a simple extrapolation (Lagrangian persistence) forecast.

2) What are the accuracy and skill of rainfall rate and occurrence forecasts as a function of lead time and accumulation period?

   The answer to this question depends not only on the inherent skill of the algorithm, but also its spatial and temporal resolution, and whether the gauge data or radar analyses are used as ‘‘truth.’’ Almost all of the algorithms gave good predictions of the overall rain frequency throughout the forecast period. Closer examination of the rain-rate distribution showed that the frequency of high rain rates was underestimated by most of the algorithms. Comparison with rain

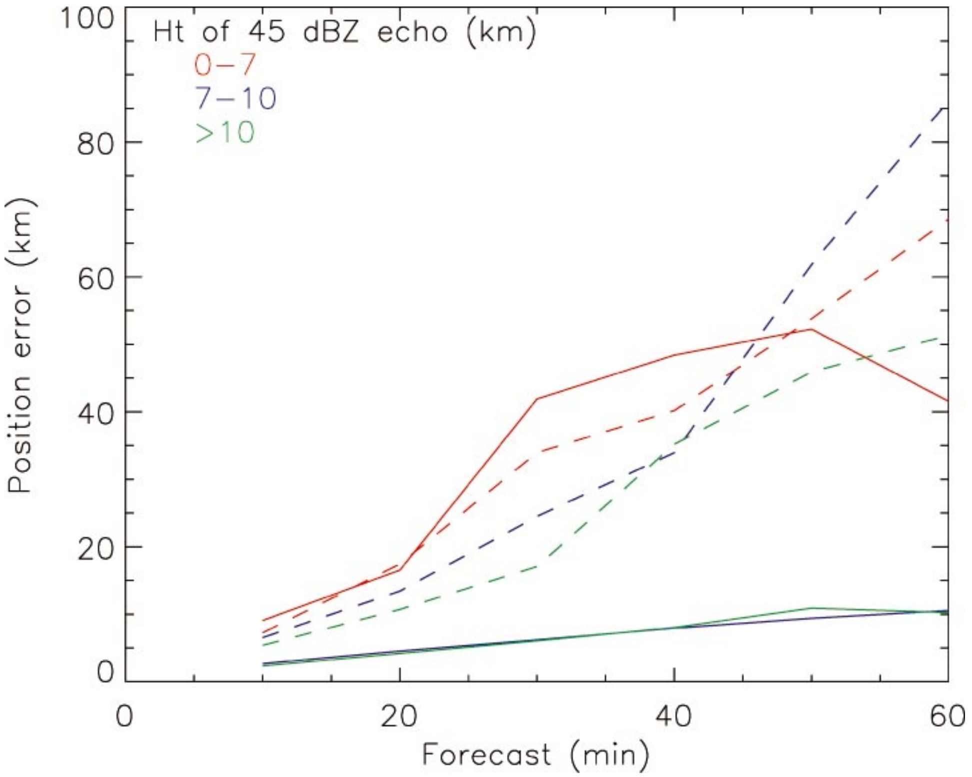

FIG. 25. Mean cell position errors for TIFS nowcasts (solid lines) and original unfiltered nowcasts (dashed lines), for three ranges of cell height.

accumulation at gauges showed that most of the algorithms were biased slightly low. This could easily be corrected by adjusting the *Z–R* relationship used to calculate the rain rates (*R*) from the radar reflectivities (*Z*).

When used to verify spatial forecasts, the CSI quantifies the ability of the algorithms to forecast rain in the correct location. The CSI values dropped very quickly with increasing forecast period, with typical values of about 0.2 after 60 min. The CSI for rain $\geq$20 mm h^{-1} was essentially nil after 30 min, indicating that the algorithms were unable to skillfully predict the precise location of heavy rain. In so far as the heaviest rain is usually embedded in a larger field of lighter rain, the nowcasts can still be considered useful for predicting the general location of heavy rain.

The algorithms consistently performed much better than radar persistence in predicting the spatial distribution of rain. When predicting rain occurrence at point locations, gauge persistence was difficult to beat, mainly because of differences in the spatial and temporal resolution. The MAE-based skill with respect to (gauge) persistence was negative early in the forecast period, then increased with time. The algorithms that performed best according to this measure were those that did not predict very much heavy rain later in the forecast period, thus avoiding the double penalty of ''rain in the wrong spot, no rain in the right spot.'' Whether this is a feature to be desired depends on the needs of the user.

3) Is it feasible to predict wind speed and direction at points with enough accuracy and skill to be useful?

This question could not be answered because the nowcast algorithms did not explicitly predict wind speed. The Auto-nowcaster scheme was an exception, but the adjoint winds produced by the mesoscale model assimilation of Doppler velocities (Sun and Crook 2001) were not considered robust enough during the FDP to be quantitatively verified.

4) What is the accuracy of severe thunderstorm wind gust diagnoses and forecasts?

The Auto-nowcaster was the only algorithm to predict the motion of gust fronts. Although the sample size was fairly limited, the verification results showed a mean absolute error of about 7 km h^{-1}, and a mean bias of 3 km h^{-1} in the speed of the gust fronts during the FDP. The errors were smaller for the prediction of sea-breeze fronts, with a mean absolute error of 3.5 km h^{-1} and bias of -0.3 km h^{-1}. The length of the fronts was also verified, with mean errors between 6 and 18 km (7%–25%) for 1-h forecasts.

5) What is the accuracy of hail location and size detections and forecasts?

It was possible to verify only one hail case, namely that associated with the tornadic storm on 3 November. The two algorithms that estimated hail size and occurrence, CARDS and WDSS, successfully diagnosed the onset and cessation of the hail to within 30 min of the reported sightings from ground-based observers. The time evolution of hail size was reasonably well captured by the algorithms, and the predicted mean and maximum hail diameters were consistent with what was observed. It would have been better to have had a more extensive dataset, but the results for the 3 November case indicate that both algorithms showed notable skill in hail detection.

6) Do the forecasters improve the quality of the forecasts compared to the ''raw'' FDP products alone?

The TIFS system allowed this question to be at least partially addressed, and the results were strongly positive. By giving forecasters the ability to modify the output of the cell tracking nowcasts, they were able to remove cells that were insignificant or diagnosed with incorrect motion. In essence, the forecasters could ''clean up'' the forecasts before sending them out to clients. About two-thirds of the cells were manually filtered out over the 5 days in which TIFS was used. The verification results showed that the mean cell position errors were markedly reduced when compared to the unfiltered forecasts, particularly for the more intense storms. However, when forecasters attempted to adjust the storm tracks for a small number of well-defined intense storms, the position errors increased by 20%, suggesting that the objective guidance is probably the best estimate of storm motion in these cases. Further testing for a much larger number of samples is necessary to more accurately assess the effects of forecaster intervention.

A related question is whether the quality of the forecasts issued to the public was enhanced as a result of the objective guidance provided by the nowcast algorithms. In the survey discussed by Anderson-Berry et al. (2004), the Sydney forecasters indicated that they would have liked more time to become comfortable with the nowcast systems prior to the FDP, and so did not make optimum use of them during the high-stress severe storm situations that they were designed to help with. However, the forecasters believed, and the verification results here strongly confirm, that the nowcast algorithms do indeed provide useful guidance for predicting severe weather, rain, and boundary layer convergence lines.

It is also important to note that most nowcast systems were designed to be used with forecasters as a supporting element. To minimize potentially damaging misses and alert the forecaster to potential worst-case scenarios, many systems inherently overforecast extreme conditions, then allow the forecaster to make the final decision. Such algorithms will generally have poorer verification results than if they had been optimized to achieve the best scores (but provide less value in an operational setting).

This verification exercise provided a unique opportunity to evaluate a variety of different nowcasting schemes in a comparable setting. Unfortunately, it was not possible to directly compare the nowcasts produced by these systems, due to the important differences in their input data, spatial and temporal resolution, and their output. Ideally, future demonstration programs will involve verification planning at the outset of the process of designing the program, so that these sources of differences can be avoided. In order to compare forecasts, it is critical to carefully control any other factors that might lead to differences in the outcomes. For example, a single, agreed-upon, ''best'' high-resolution radar reflectivity field using optimized quality control, calibration, etc. should be available for initializing and verifying the algorithms.

This exercise also provides an impetus for advancing the science of verification through development of more meaningful verification approaches that are appropriate for evaluating mesoscale forecasts (e.g., Ebert and McBride 2000). Although the statistics presented here do provide measures of the overall quality of the forecasts, and they indicate some strengths and weaknesses in the systems, in general they do not provide specific guidance regarding which aspects of the forecasts need to be repaired, or how those repairs should be done. In-depth evaluations of the individual algorithms should involve more complete characterization of a variety of types of errors associated with the forecasts (e.g., location, size, timing, etc.), which could facilitate these types of diagnoses.

Acknowledgments. The authors would like to thank Alan Seed for providing the motion vectors used to make the extrapolation forecasts, and Randy Bullock for his assistance in evaluating the TITAN location errors. This paper benefited from helpful discussions with Peter May, Rod Potts, Clive Pierce, and especially Jim Wilson, as well as comments from the anonymous reviewers.

REFERENCES

Anderson-Berry, L., T. Keenan, J. Bally, R. Pielke Jr., R. Leigh, and D. King, 2004: The societal, social, and economic impacts of the World Weather Research Programme Sydney 2000 Forecast Demonstration Project (WWRP S2000 FDP). *Wea. Forecasting,* **19,** 168–178.

Bally, J., 2004: The Thunderstorm Interactive Forecast System: Turning automated thunderstorm tracks into severe weather warnings. *Wea. Forecasting,* **19,** 64–72.

Bellon, A., and G. L. Austin, 1978: The evaluation of two years of real-time operation of a short-term precipitation forecasting procedure (SHARP). *J. Appl. Meteor.,* **17,** 1778–1787.

Brown, B. G., and E. Brandes, 1997: An intercomparison of 2D storm motion extrapolation algorithms. Preprints, *28th Int. Conf. on Radar Meteorology,* Austin, TX, Amer. Meteor. Soc., 495–496.

——, G. Thompson, R. T. Bruintjes, R. Bullock, and T. Kane, 1997:

Intercomparison of in-flight icing algorithms. Part II: Statistical verification results. *Wea. Forecasting,* **12,** 890–914.

——, and Coauthors, 2001: Forecast verification activities for the Sydney 2000 Forecast Demonstration Project. Preprints, *30th Int. Conf. on Radar Meteorology,* Munich, Germany, Amer. Meteor. Soc., 500–502.

Dixon, M., and G. Wiener, 1993: TITAN: Thunderstorm Identification, Tracking, Analysis, and Nowcasting—A radar-based methodology. *J. Atmos. Oceanic Technol.,* **10,** 785–797.

Donaldson, N., C. Pierce, M. Sleigh, A. Seed, and T. Saxen, 2001: Comparison of forecasts of widespread precipitation during the Sydney 2000 Forecast Project. Preprints, *30th Int. Conf. on Radar Meteorology,* Munich, Germany, Amer. Meteor. Soc., 503–505.

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

Eilts, M., and Coauthors, 1996: Severe weather warning decision support system. Preprints, *18th Conf. on Severe Local Storms,* San Francisco, CA, Amer. Meteor. Soc., 536–540.

Golding, B. W., 1998: Nimrod: A system for generating automated very short range forecasts. *Meteor. Appl.,* **5,** 1–16.

Johnson, J. T., P. L. MacKeen, A. Witt, E. D. Mitchell, G. J. Stumpf, M. D. Eilts, and K. W. Thomas, 1998: The Storm Cell Identification and Tracking algorithm: An enhanced WSR-88D algorithm. *Wea. Forecasting,* **13,** 263–276.

Keenan, T., 1999: Hydrometeor classification with a C-band polarimetric radar. Preprints, *29th Int. Conf. on Radar Meteorology,* Montreal, QC, Canada, Amer. Meteor. Soc., 184–187.

——, and Coauthors, 2003: The Sydney 2000 World Weather Research Program Forecast Demonstration Project: Overview and current status. *Bull. Amer. Meteor. Soc.,* **84,** 1041–1054.

Lapczak, S., and Coauthors, 1999: The Canadian National Radar Project. Preprints, *29th Int. Conf. on Radar Meteorology,* Montreal, QC, Canada, Amer. Meteor. Soc., 327–330.

Mason, S. J., and G. M. Mimmack, 1992: The use of bootstrap confidence intervals for the correlation coefficient in climatology. *Theor. Appl. Climatol.,* **45,** 229–233.

Nurmi, P., J. Bally, H. E. Brooks, B. G. Brown, E. Ebert, M. Jaeneke, and L. Wilson, 2001: International cooperation in developing advanced nowcasting tools—The verification study of the Sydney 2000 Forecast Demonstration Project. *Extended Abstracts, Fifth European Conf. on Applications of Meteorology,* Budapest, Hungary, European Meteorology Society, CD-ROM.

Pierce, C. E., C. G. Collier, P. J. Hardaker, and C. M. Haggett, 2000: GANDOLF: A system for generating automated nowcasts of convective precipitation. *Meteor. Appl.,* **7,** 341–360.

——, and Coauthors, 2004: The nowcasting of precipitation during Sydney 2000: An appraisal of the QPF algorithms. *Wea. Forecasting,* **19,** 7–21.

Seed, A., and T. Keenan, 2001: A dynamic and spatial scaling approach to advection forecasting. Preprints, *30th Int. Conf. on Radar Meteorology,* Munich, Germany, Amer. Meteor. Soc., 492–494.

Sun, J., and N. A. Crook, 2001: Real-time low-level wind and temperature analysis using WSR-88D data. *Wea. Forecasting,* **16,** 117–132.

Webb, R. M., A. Treloar, J. Colquhoun, R. Potts, J. Bally, T. Keenan, and P. May, 2001: Overview of Sydney weather during the Forecast Demonstration Project. Preprints, *30th Int. Conf. on Radar Meteorology,* Munich, Germany, Amer. Meteor. Soc., 477–479.

Wilson, J. W., N. A. Crook, C. K. Mueller, J. Sun, and M. Dixon, 1998: Nowcasting thunderstorms: A status report. *Bull. Amer. Meteor. Soc.,* **79,** 2079–2099.

Witt, A., M. D. Eilts, G. J. Stumpf, J. T. Johnson, E. D. Mitchell, and K. W. Thomas, 1998: An enhanced hail detection algorithm for the WSR-88D. *Wea. Forecasting,* **13,** 286–303.