

Verification of Supercooled Cloud Water Forecasts with In Situ Aircraft Measurements

HONG GUAN, STEWART G. COBER, AND GEORGE A. ISAAC

Cloud Physics Research Division, Meteorological Service of Canada, Downsview, Ontario, Canada

(Manuscript received 22 December 1999, in final form 15 June 2000)

ABSTRACT

In situ measurements of temperature (T_a), horizontal wind speed (V), dewpoint (T_d), total water content (TWC), and cloud and supercooled cloud water (SCW) events, made during 50 flights from three research field programs, have been compared to forecasts made with the High Resolution Model Application Project version of the Global Environmental Multiscale model. The main purpose of the comparisons was to test the accuracy of the forecasts of cloud and SCW fields. The forecast accuracy for T_a , V , and T_d agreed closely with the results from radiosonde–model validation experiments, implying that the aircraft–model validation methodology was equally feasible and, therefore, potentially applicable to SCW forecast verifications (which the radiosondes could not validate).

The hit rate (HR), false alarm rate (FAR), and true skill statistic (TSS) for cloud forecasts were found to be 0.52, 0.30, and 0.22, respectively, when the model data were inferred at a horizontal resolution of 1.5 km (averaging scale of the aircraft data). The corresponding values for SCW forecasts were 0.37, 0.22, and 0.15, respectively. The HRs (FARs) for cloud and SCW events are sensitive to horizontal resolution and increase to 0.76 (0.50) and 0.66 (0.53), respectively, when a horizontal resolution of 100 km is used. The model TWC was found to agree poorly with aircraft measurements, with the model generally underestimating TWC. For cases when the forecasts and observations of cloud agreed, the SCW-forecast HR, FAR, and TSS were 0.63, 0.22, and 0.41, respectively, which implies that improvement in the model cloud fields would substantially improve the SCW forecast accuracy.

The demonstrated comparison methodology will allow a quantitative comparison between different SCW and cloud algorithms. Such a comparison will provide insight into the strengths and weaknesses of these algorithms and will allow the development of more accurate cloud and SCW forecasts.

1. Introduction

In-flight aircraft icing continues to cause aviation accidents. In Canada between 1980 and 1990, there were 937 aircraft accidents with 585 fatalities, in which weather-related phenomena were recognized as a contributing factor in the accident. Of these, 52 accidents with 273 fatalities had icing as a recognized contributing factor. Clearly, avoidance of icing regions through accurate forecasts of supercooled cloud water is an important research objective for the aviation and forecasting communities. This was the motivation for this study.

A number of numerical diagnostic aircraft icing algorithms have been developed and evaluated (Appleman 1954; Knapp 1992; Schultz and Politovich 1992; Forbes et al. 1993; Tremblay et al. 1995, 1996). Most of these algorithms test whether the model output at a grid point

meets specific temperature, humidity, or vertical velocity criteria, although the Tremblay algorithm is based on parameterizations of physical processes. Tremblay et al. (1996) demonstrated that their forecast procedures produced a decreasing distribution of icing frequency as temperature decreased, which is in agreement with aircraft observations (Cober et al. 1995) and climatology studies of pilot-report icing (Rasmussen et al. 1992).

In this paper, the Canadian Global Environmental Multiscale (GEM) model derived temperature (T_a), dewpoint (T_d), horizontal wind speed (V), and cloud total water content (TWC) will be validated in a four-dimensional framework by direct comparison with in situ aircraft measurements, and the results will be compared to model–radiosonde validation results. Similarly, the Sundqvist cloud scheme (Sundqvist et al. 1989) and the Tremblay supercooled cloud water (SCW) forecasting scheme will be compared with the in situ aircraft data by comparing the observed or not observed aircraft measurements with the forecast or not forecast model results. The dataset is large enough to allow computation of hit rate (HR), false alarm rate (FAR), and true skill statistic (TSS) for the cloud and SCW forecasting al-

Corresponding author address: Hong Guan, Cloud Physics Research Division, Meteorological Service of Canada, 4905 Dufferin Street, Downsview, ON M3H 5T4, Canada.
E-mail: Hong.Guan@ec.gc.ca

gorithms. The GEM model, Sundqvist cloud scheme, and Tremblay SCW scheme were selected because they are currently used for the Canadian operational aviation forecast. The relative skills of the Appleman and Tremblay SCW forecasting schemes will also be assessed. This is significant because the Tremblay scheme replaced the Appleman scheme in 1994 in the Canadian operational aviation forecast. The Appleman scheme is still used operationally in the automated icing forecast program of the U.S. Air Force Global Weather Center. It is hoped that this work will demonstrate a useful validation tool for developers of numerical forecast algorithms of cloud and SCW.

The aircraft data were collected during 50 research flights in three field programs. The field projects include the Second Canadian Atlantic Storms Program (CASP II), and the First and Third Canadian Freezing Drizzle Experiments (CFDE I and III, respectively), all of which were designed in part to measure and characterize aircraft icing regions in winter storms. Project flights covered a wide variety of different meteorological conditions and geographical locations. CASP II and CFDE I were based in St. John's, Newfoundland, and hence focused on maritime winter storms, while CFDE III was based in Ottawa, Ontario, and focused on continental winter storms.

2. Model algorithms

The numerical model used was the experimental High Resolution Model Application Project (HIMAP) version of the GEM model (Côté et al. 1998a,b), with 0.14° horizontal grid spacing (corresponding to approximately 15 km in the middle of the domain) and 35 eta levels. Two different uniform-resolution subdomains were used to ensure the research aircraft flight tracks were located near the centers of the subdomains. HIMAP East, which is centered on the Quebec–Windsor corridor, was used for CFDE III simulations. For CASP II and CFDE I simulations, the HIMAP grid was centered on St. John's, with the same horizontal domain size (200 × 251 grid points). The meteorological fields used to initialize the model were obtained from the operational Canadian Meteorological Centre (CMC) objective analysis at 0600 UTC. Similar to the CMC operational GEM (HIMAP version) run, the model integrates from 0600 UTC for every flight. The validation periods of the forecasts ranged from 4 to 25 h, depending on the flight time. Figure 1 shows the distribution of model validation times in 6-h intervals. The majority of the validation periods were between 9 and 15 h from initialization. No attempts were made to differentiate model forecast accuracies for different forecast times. The physical fields (such as T_a , humidity, vertical velocity, etc.) that were derived by the model, and required as input for the SCW forecasting algorithms, were generated and saved hourly.

The Sundqvist explicit condensation scheme (Sundqvist et al. 1989) is used for determining TWC and cloud

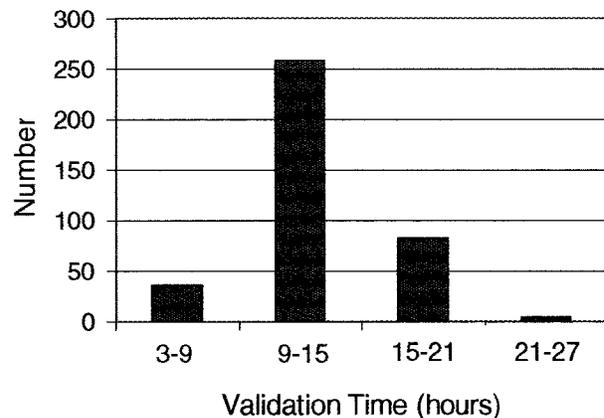


FIG. 1. Distribution of model validation times in 6-h intervals. Each observation represents 1 h during which aircraft data were collected.

forecasts. TWC (ice + water) is predicted from a continuity equation, and there is no distinction between liquid and solid phases. The scheme predicts TWC and cloud fraction, with cloud fraction being parameterized as a function of relative humidity. These parameters are also input into the Tremblay SCW forecast algorithm.

In the Tremblay SCW scheme (Tremblay et al. 1995), SCW events are inferred based on the following criteria: T_a must be between -15.5° and 0°C , TWC must be higher than a threshold value of 0.05 g m^{-3} , and $wG - \text{SDEP} > 0$, where w is vertical velocity, G is the vertical gradient of the saturation mixing ratio, and SDEP is the vapor deposition on snow. These variables are determined explicitly in the model. The last criterion physically represents microphysical processes involved in the production of supercooled cloud water, and indicates that if the amount of vapor condensed by wet-adiabatic cooling exceeds the rate of vapor deposition on snow, SCW will be produced. The criteria for the Appleman SCW forecasting scheme are T_a between 0° and -16°C , upward vertical motion greater than 0.2 Pa s^{-1} , and $T_d - T_a$ must be less than 0.2 Pa .

3. Aircraft data

Measurement campaigns have been conducted in Newfoundland (1992, 1995) and Ottawa (1997–98) using the National Research Council Convair-580 research aircraft. CASP II was conducted out of St. John's between 15 January 1992 and 15 March 1992. CFDE I was conducted out of St. John's during March 1995. CFDE III was based out of Ottawa from December 1997 to February 1998. This study used data from 27 flights from CASP II, 12 from CFDE I, and 11 from CFDE III. It is recognized that the dataset is limited in terms of the geographical regions represented and the short durations of the field projects, although climatological studies have indicated that Newfoundland and the Great Lakes area are the areas with the highest frequency of freezing precipitation in North America (Isaac et al. 1999). The research aircraft

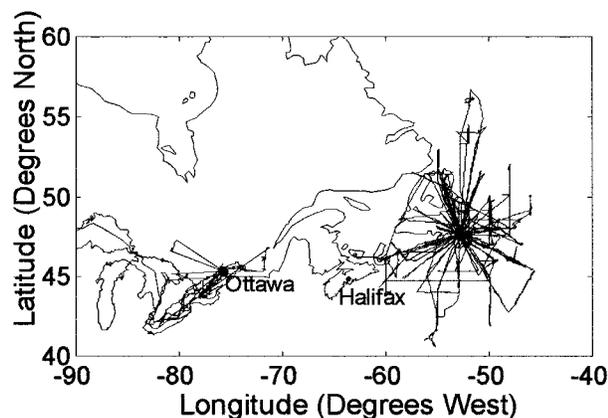


FIG. 2. Summary of flight tracks for CASP II (27 flights), CFDE I (12 flights), and CFDE III (11 flights) for data used in this study. The flights originated from St. John's, Newfoundland, or Ottawa, Ontario.

was generally directed into regions where SCW was forecast to exist, and when SCW was found, the aircraft was used to make extensive vertical and horizontal profiles in order to obtain sufficient data to allow some understanding of the SCW formation mechanisms. Considering the large number of storms measured during the field projects, the dataset is believed to be highly suited for validation of SCW forecast algorithms.

The Convair-580 was fully equipped for cloud microphysics measurements, and the instrumentation used during CASP II, CFDE I, and CFDE III are described by Cober et al. (1995, 1999) and Isaac et al. (1998). Measured parameters that were applicable to this study were V , T_a , T_d , TWC (only for CFDE I and CFDE III), liquid water content (LWC), ice crystal concentration, pressure, latitude, and longitude. The value of T_a was measured by two Rosemount temperature probes and a reverse-flow temperature probe and was accurate to $\pm 1^\circ\text{C}$; T_d was measured to within $\pm 2^\circ\text{C}$ with a Cambridge Dew-point Hydrometer. Latitude and longitude were measured with the Global Positioning System, and with inertial and loran navigation systems, and were accurate to within ± 100 m. The aircraft altitude was measured with respect to pressure, with an accuracy of 1 mb. The value of V was measured within ± 1 m s^{-1} with a Rosemount 858 probe. LWC was measured with Commonwealth Scientific and Industrial Research Organisation King and Nevzorov LWC probes to within ± 0.02 g m^{-3} , while TWC was measured to within ± 0.02 g m^{-3} with a Nevzorov TWC probe (Korolev et al. 1998).

Each flight lasted between 3 and 6 h with the average flight being approximately 4 h in duration. The aircraft data were averaged over 15 s, representing a horizontal distance of approximately 1.5 km. A summary of the flight tracks that were used in this study is shown in Fig. 2. The 50 flights selected were chosen from a larger collection of 83 flights. Flights were selected based on successful measurements of SCW, and hence there may be a bias toward synoptic meteorological cases with

TABLE 1. The contingency table used to evaluate forecasts of cloud and SCW.

	Yes forecast	No forecast	Total
Yes observed	A	B	A + B
No observed	C	D	C + D

stronger SCW signals. This may cause the HR and FAR values to be optimistic since the model physics would be expected to be more accurate in such cases.

4. Comparison methodology

To compare model forecasts and aircraft measurements, the model fields at 15-km and 1-h resolution were numerically interpolated to 1.5-km resolution along the 3D-time aircraft trajectories. The time series of T_a , T_d , V , and TWC were then quantitatively compared with 15-s (1.5 km) averaged aircraft measurements. Scatterplots of model versus aircraft measurements were used to evaluate the rms errors of the model parameters. A similar comparison methodology is used when model and radiosonde measurements are compared.

A different comparison technique was used to assess the cloud and SCW forecasts. Model cloud and SCW fields were assessed as forecast or not forecast, while aircraft cloud and SCW fields were assessed as observed or not observed. This required defining thresholds for some of the model and aircraft parameters.

The existence of observed and forecasted clouds was inferred if the TWC was higher than 0.03 g m^{-3} . However, a different criterion was used for CASP II observations because of the lack of a TWC measurement. In these cases, cloud was inferred for either an $LWC \geq 0.03$ g m^{-3} , or an ice crystal concentration ≥ 1 L^{-1} (Cober et al. 1995), as measured by a Particle Measuring System 2DP probe (200–6400 μm). The 2DP was not working on 9 of the 27 CASP II flights, so that the presence of cloud where no liquid water was present was based on a comparison between T_a and T_d . Cloud was inferred if the difference was within 2°C . While the model SCW field was output as a yes/no value, the aircraft measurements were inferred to indicate SCW events when the average supercooled liquid water content (SLWC) exceeded 0.04 g m^{-3} . Sensitivity tests demonstrated that the use of SCW thresholds between 0.03 and 0.1 g m^{-3} only produced small changes in the HR, FAR, and TSS results.

Using these criteria, for each 1.5-km data point with coincident model and aircraft data, two sets of yes/no observations for the SCW and cloud fields were assessed. Table 1 incorporates these observations into the basic 2×2 contingency table associated with dichotomous forecasts. The quality of the cloud and SCW forecasts was evaluated based on signal detection theory (SDT; Mason 1982) and analysis of the true skill statistics (Flueck 1987; Wilks 1995). These scores, calculated from the 2×2 contingency table, are applied

widely in weather forecast evaluations (Swets 1988; Dossell et al. 1990; Tremblay et al. 1996).

In SDT, the forecast skill is related to the HR and FAR. Following Table 1, the HR $[A/(A + B)]$ can be interpreted as the proportion of observed cloud or SCW events that were correctly forecast, while the FAR $[C/(C + D)]$ is the proportion of not-observed cloud or SCW events that were forecast. These are equivalent to the probability of detection and probability of false detection parameters described by Tremblay et al. (1996). The TSS is

$$\text{TSS} = \text{HR} - \text{FAR} = \frac{(AD - BC)}{(A + B)(C + D)}, \quad (1)$$

where $-1 \leq \text{TSS} \leq 1$, with 1 representing a perfect forecast. A value of 0 represents a forecast with no skill.

Initially, the model forecast fields interpolated to 1.5-km resolution were directly compared with the averaged aircraft data at equal positions and times. In this situation, the aircraft data averaged over a specific 1.5-km track were compared with data from a single 1.5-km model grid point. However, a forecast model can produce spatial error as discussed in Moninger et al. (1991) and Tremblay et al. (1996). To estimate the sensitivity of the verification results to spatial precision, the model fields were assessed over grids of successively larger squares centered on each individual aircraft-based measurement. Since the model data were known for each $1.5 \text{ km} \times 1.5 \text{ km}$ grid point, this required comparing each aircraft data point with several model grid points that surrounded and overlapped the aircraft point. For example, comparison of a 1.5-km aircraft data point with a $16 \text{ km} \times 16 \text{ km}$ model field would require comparison with the 121 model grid points (each, $1.5 \text{ km} \times 1.5 \text{ km}$) that make up the single $16 \text{ km} \times 16 \text{ km}$ grid field. Model fields of $1, 11 \times 11, 17 \times 17, 33 \times 33,$ and 67×67 grid points, representing boxes of length 1.5, 16, 25, 50, and 100 km, respectively, were compared to the 1.5-km aircraft measurements. If a forecast value at any of the model grid points was yes, then a yes forecast was assigned to the model field. Conversely, if all of the grid point forecasts were no, the model field was assigned a no forecast. The sensitivity of the results to different yes/no thresholds is discussed in section 5. The 50- and 100-km comparisons were performed because these scales are also useful to the users of aviation forecasts of cloud and SCW. The 16- and 25-km comparisons were performed because these scales are close to the standard grid scales of the last two versions of the GEM model. The effect of vertical errors from the forecast system on the verification results was not tested.

5. Verification results

a. Validation of meteorological fields

The aircraft–model comparisons for Ta, Td, and V were done to test the comparison methodology. Figure

3 shows comparisons of the aircraft and model data for the CASP II, CFDE I, and CFDE III projects. The corresponding rms errors are also shown. The rms errors for Ta were similar for all three projects with a minimum error of 1.7°C for CFDE III. Larger scatter is indicated by the rms errors for Td and V. The CFDE III data consistently showed the smallest rms errors for Ta, Td, and V in comparison to the CASP II and CFDE I data. This may have been related to the higher density of surface observations (i.e., model initialization data) for the CFDE III cases. Unfortunately, the dataset is not large enough to quantitatively support this conclusion.

The Ta, V, and dewpoint depression forecast accuracies agreed closely with the results from radiosonde–model validation experiments reported by Côté et al. (1998a). It should be noted that the model data for the current study were based on the GEM model at 15-km resolution, which is different than the 35-km resolution version used in Côté et al. (1998a). However, a radiosonde–model validation using the 15-km resolution version gave very similar results to those given in Côté et al. (1998a) (A. Methot 1999, personal communication). Figure 4 shows a comparison between the aircraft–model validation results and the Côté et al. (1998a) radiosonde–model validation results, the latter based on 120 North American stations. Their rms error profiles were derived for 12-h forecasts, comparable with the average validation time for the results presented here. The results indicate that the overall performances for both sets of validations are very similar. The rms errors for V and dewpoint depression for CFDE III are slightly smaller than those in the radiosonde–model validation. Conversely, the rms errors for Ta and V for CASP II and CFDE I are slightly larger than the radiosonde–model results.

The consistency of the aircraft–model comparison with the radiosonde–model comparison of Côté et al. (1998a) demonstrates that the aircraft–model comparison methodology is an equally feasible approach to model validation. Both techniques will have some error associated with trying to compare data averaged at a point or along a line with data averaged over a volume. The aircraft–model comparison technique allows the opportunity to evaluate the model accuracy for parameters such as LWC and TWC, which are not measured with the radiosondes. It also allows a quantitative technique for intercomparing different SCW algorithms. This is particularly beneficial considering that the aircraft data were obtained in meteorological conditions that generally had a strong SCW signal.

b. Validation of cloud forecasts

The cloud verification statistics, based on the use of single grid points ($1.5 \text{ km} \times 1.5 \text{ km}$), for CASP II, CFDE I, and CFDE III are shown in Table 2. There were similar HR values for all three projects; however, the FAR and TSS values showed some differences. CFDE III had the lowest FAR (0.24) and therefore the

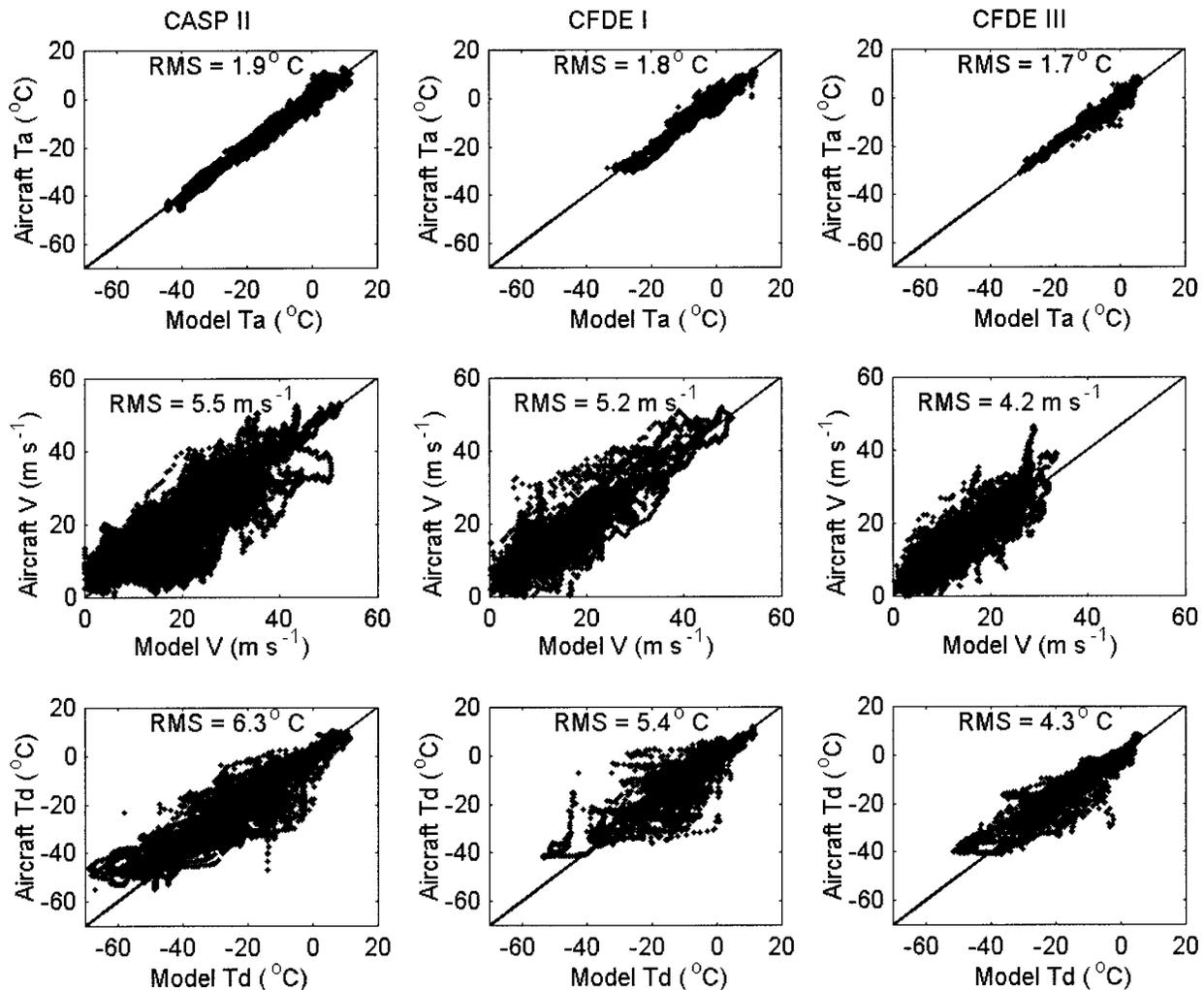


FIG. 3. Comparisons of the measured and forecast temperature (T_a), horizontal wind speed (V), and dewpoint (T_d) for CASP II, CFDE I, and CFDE III. The solid curves represent a 1:1 correlation.

highest TSS value (0.27), while CFDE I had the highest FAR (0.36) and the lowest TSS value (0.18).

The effect of spatial error on the model forecast was estimated by evaluating the HR, FAR, and TSS over successively larger model areas as discussed in section 4. The results of this sensitivity study are given in Table 3. The category labeled Total represents the analysis for the collective dataset. There is an expected trade-off between HR and FAR pairs, with the HR and FAR increasing with increasing model grid size. Overall, the HR (FAR) increased from 0.52 (0.30) at $1.5 \text{ km} \times 1.5 \text{ km}$ resolution to 0.76 (0.50) at $100 \text{ km} \times 100 \text{ km}$ resolution. The corresponding TSS increased from 0.22 to 0.26, indicating a slightly higher skill at the larger model grid scale. To make a comparison with the CMC two-dimensional verification result for the total cloud opacity, the Heidke skill score (Panofsky and Brier 1965) was also calculated. The values ranged from 0.22 to 0.25, and are smaller than the 0.4 value (CMC 1998)

determined by CMC from a total column cloud opacity validation. The additional forecast error of cloud in the vertical may partly account for this difference. The TSS value is well above the useless forecast value, indicating that the forecast model has some ability to forecast cloud, although there is room for considerable improvement. An improvement in the model humidity field would undoubtedly lead to better cloud forecasts.

Using a single yes observation from a 1.5-km grid point to define a yes observation for a model field that has, for example, 121 grid points (i.e., $16 \text{ km} \times 16 \text{ km}$) could lead to an unreasonable bias in the HR and FAR statistics. For the 67×67 grid point field, a yes forecast would be scored even if there were 4488 no and 1 yes observations. To determine the sensitivity of the results to this criterion, the HR, FAR, and TSS were recalculated using different thresholds for a yes forecast. A yes forecast was inferred only when the number of yes forecasts from the 1.5-km grids exceeded some percentage

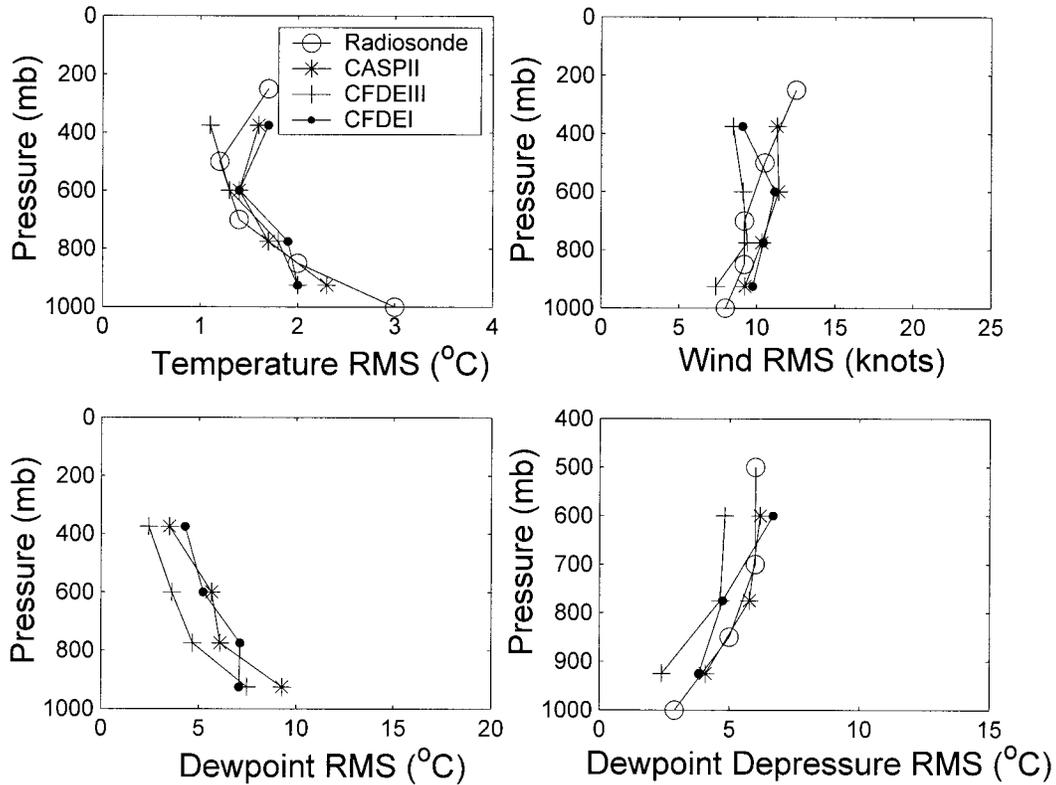


FIG. 4. Vertical rms errors for Ta, V, Td, and dewpoint depression for CASP II, CFDE I, and CFDE III. Radiosonde indicates the results from the model-radiosonde comparisons.

threshold between 10% and 50% of the total number of 1.5-km grid points. Table 4 shows the changes in HR, FAR, and TSS for all flights for thresholds of one pixel, 10%, 25%, and 50%. As expected, there is a small decrease in HR and FAR as the threshold is increased from one pixel to 50% of the pixels. The magnitude of the difference increases with increasing grid size. For the worst case (the 67 × 67 grid comparison) the HR (FAR) change from 0.76 (0.50) to 0.52 (0.30) as the threshold changes from one pixel to 50%. However the TSS changes very little, which is true for all cases considered. Therefore, the forecast skill is not significantly affected by the yes/no methodology chosen. For the 11

TABLE 2. Cloud forecast verifications for 1 × 1 model grid points. The data represent a resolution of 1.5 km.

Project	Ob-served	Forecast		Total	Statistics
		Yes	No		
CASP II	Yes	5873	5476	11 349	HR = 0.52
	No	4762	10 715	15 477	FAR = 0.31
					TSS = 0.21
CFDE I	Yes	2956	2574	5530	HR = 0.53
	No	1991	35841	5572	FAR = 0.36
					TSS = 0.18
CFDE III	Yes	2616	2523	5139	HR = 0.51
	No	1419	4573	5992	FAR = 0.24
					TSS = 0.27

× 11 and 17 × 17 grid comparisons the changes in HR, FAR, and TSS are essentially independent of the threshold used to infer a yes forecast. Considering that cloud fields in winter storms tend to be quite large in hori-

TABLE 3. Cloud forecast verifications using multiple grid points for model data. The data represent resolutions of 1.5, 16, 25, 50, and 100 km.

Project	Grid points	HR	FAR	TSS
CASP II	1 × 1	0.52	0.31	0.21
	11 × 11	0.56	0.35	0.21
	17 × 17	0.59	0.37	0.22
	33 × 33	0.66	0.43	0.23
	67 × 67	0.74	0.48	0.26
CFDE I	1 × 1	0.53	0.36	0.18
	11 × 11	0.59	0.40	0.19
	17 × 17	0.63	0.42	0.20
	33 × 33	0.68	0.44	0.24
	67 × 67	0.79	0.57	0.22
CFDE III	1 × 1	0.51	0.24	0.27
	11 × 11	0.56	0.28	0.28
	17 × 17	0.59	0.30	0.29
	33 × 33	0.63	0.35	0.28
	67 × 67	0.74	0.46	0.28
Total	1 × 1	0.52	0.30	0.22
	11 × 11	0.57	0.34	0.23
	17 × 17	0.60	0.37	0.23
	33 × 33	0.66	0.41	0.25
	67 × 67	0.76	0.50	0.26

TABLE 4. Cloud forecast verifications for yes forecast thresholds of one pixel, and 10%, 25%, and 50% of the total pixels.

Grid points	Threshold	HR	FAR	TSS
11 × 11	One pixel	0.57	0.34	0.23
	10%	0.55	0.33	0.22
	25%	0.54	0.32	0.22
	50%	0.52	0.30	0.22
17 × 17	One pixel	0.60	0.37	0.23
	10%	0.57	0.35	0.23
	25%	0.55	0.33	0.23
	50%	0.52	0.30	0.22
33 × 33	One pixel	0.66	0.41	0.24
	10%	0.61	0.38	0.23
	25%	0.57	0.35	0.22
	50%	0.51	0.30	0.21
67 × 67	One pixel	0.76	0.50	0.25
	10%	0.67	0.43	0.24
	25%	0.62	0.38	0.24
	50%	0.52	0.30	0.22

zontal extent, and that the research flights in this study were generally directed into regions where the clouds had a large horizontal extent, most of the data collected were in clear air regions not close to cloud (i.e., transit to the region of interest) or cloud regions not close to clear air (i.e., the region of interest) and relatively less data was collected near the boundary of the two. Hence the result described above is not surprising.

The vertical distributions of HR, FAR, and TSS for cloud forecasts at 1.5-km resolution are shown in Fig. 5. The maximum HR is located at the middle level between 700 and 800 mb, while the values decrease toward the high and low levels. The FAR was relatively constant (about 0.4) throughout the entire low and middle levels with a sharp decrease at high levels. The TSSs at the middle and high levels were much better than those at the low level, indicating a poor forecast accuracy for low-level cloud. This observation is consistent with the results of Yu et al. (1997) and suggests a model weakness in resolving boundary layer clouds.

A scatterplot of model and aircraft TWC, for CFDE I and III data, is displayed in Fig. 6. CASP II data were not included because there was no TWC measurement in this project. Only data with a TWC > 0.01 g m⁻³ (both aircraft and model) are shown in Fig. 6. The aircraft measurements were separated into liquid, mixed, and glaciated categories following techniques described in Cober et al. (1999). Liquid phase implies insignificant ice while glaciated phase implies insignificant liquid. A poor correlation is obvious for all three situations. The TWC rms errors are between 0.1 and 0.2 g m⁻³ with slightly smaller values for CFDE III. There are no obvious differences between the model-aircraft validations for the liquid, mixed, and glaciated clouds. It is evident that the model generally underestimates the mixed and glaciated cloud TWC. The implication is that the microphysics algorithms used in the Sundqvist cloud scheme do not infer cloud TWC accurately. This will

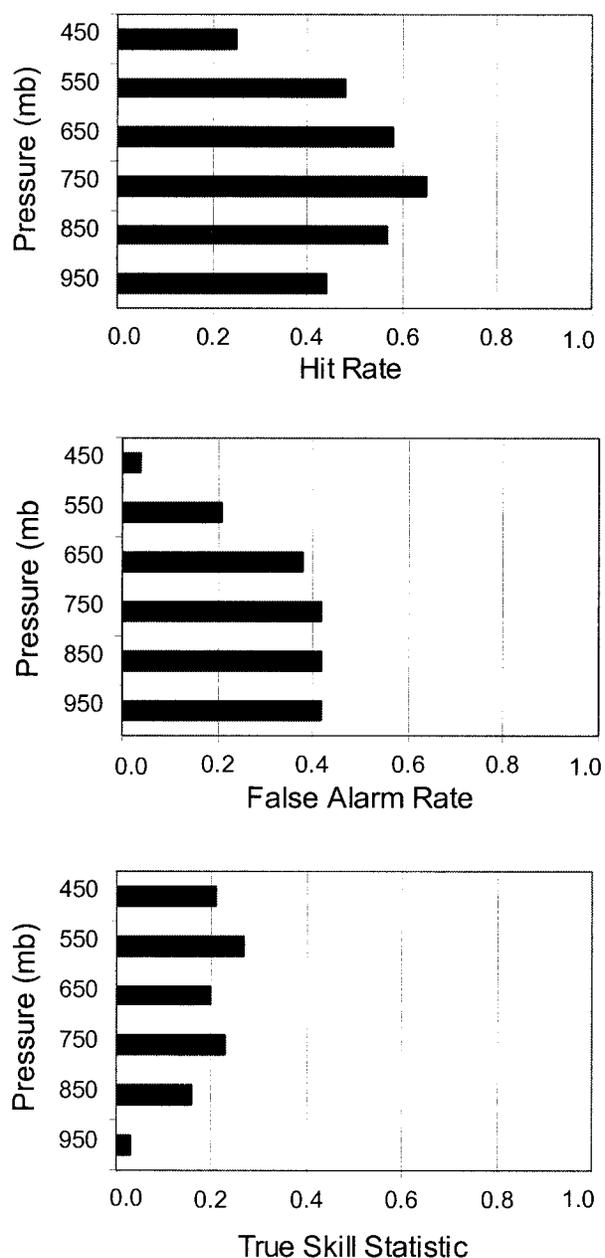


FIG. 5. Vertical distributions of HR, FAR, and TSS for the Sundqvist cloud algorithm.

have a direct effect on the SCW forecast, since the SCW forecast relies on the cloud TWC.

One possible explanation of the poor correlation between the model and aircraft measurements is that the coarse model resolution does not resolve the subgrid variability. Comparing aircraft data averaged along a line with model data averaged over a volume will add scatter to the comparison. A comparison of aircraft data averaged at 15-km resolution with model data at 15-km resolution (not shown) showed very similar results to Fig. 6. In addition, the cloud regions measured with the

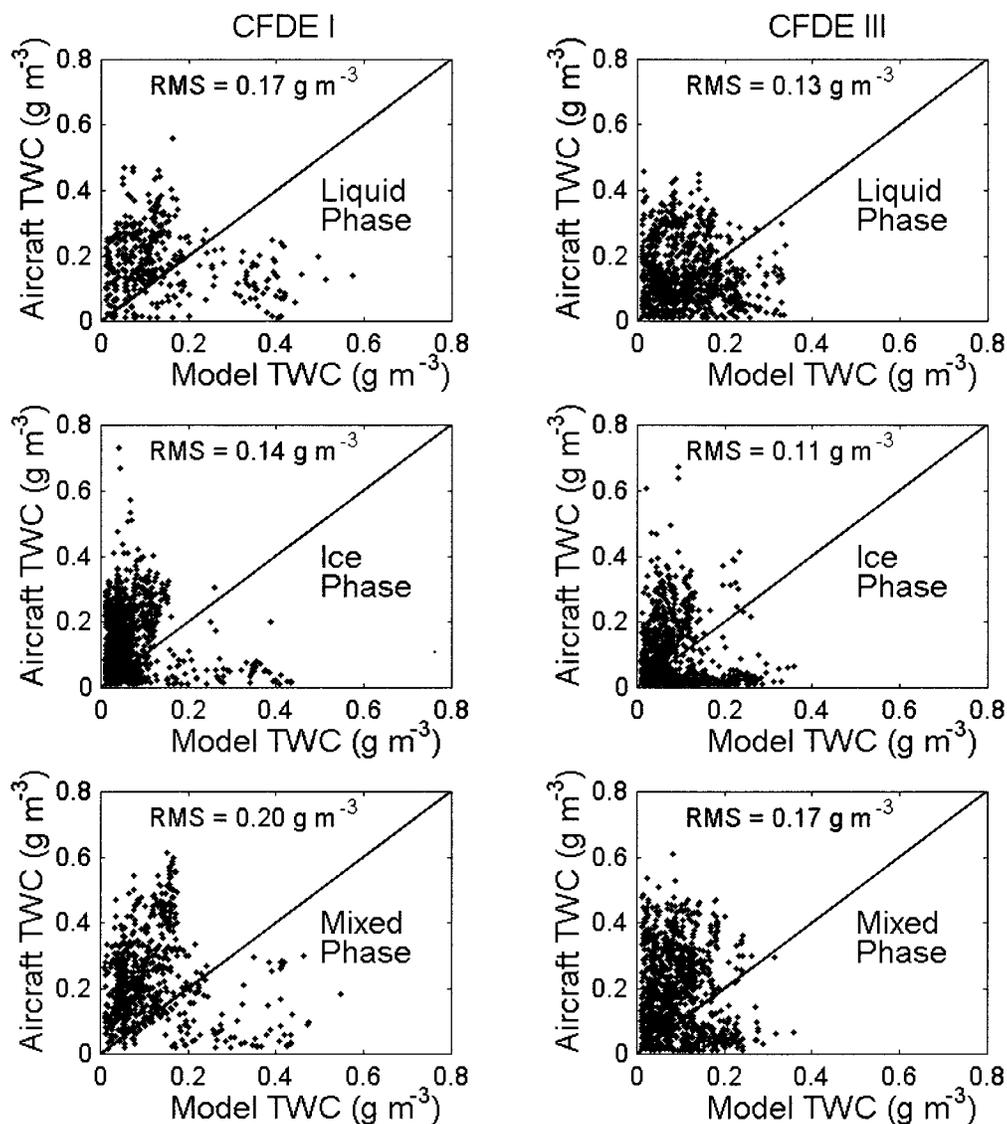


FIG. 6. Comparisons of the measured and forecast TWC for CFDE I and CFDE III. The observed cloud data have been segregated into liquid, mixed, and glaciated phase conditions. The solid curves represent a 1:1 correlation. The data represent a resolution of 1.5 km.

aircraft were generally stratiform in nature, covering a wide horizontal region. Therefore, it is not believed that the model resolution is the dominant reason for the disagreement between the aircraft and model data. Another possible explanation is that forecast errors inherent in the model dynamics cause errors in the cloud fields, which lead to the poor correlation between the model and aircraft measurements. Considering that the average validation time was 9–15 h, this is certainly a reasonable explanation for some of the scatter seen in Fig. 6.

c. Validation of SCW forecasts

The verification results for the Tremblay and Appelman SCW forecasting schemes are listed in Table 5. The

HR, FAR, and TSS values are very similar for the Tremblay scheme for CFDE I and CFDE III, while the CASP II results reflect a significantly poorer performance. The best HR and TSS values are for CFDE I, being 0.40 and 0.17, respectively. The TSS value is above the useless forecast value for CFDE I and CFDE III, although the HR is low. The TSS for CASP II is significantly lower than those for CFDE I and CFDE III, which may be caused by the small SCW dataset for CASP II. The clouds sampled during CASP II were frequently glaciated, often being associated with deep warm frontal or low pressure region clouds that had cloud-top temperatures colder than -20°C . In these clouds, regions of SCW that might be forecast by the model could be glaciated by ice particles falling from higher layers. This would cause a higher

TABLE 5. SCW forecast verifications for 1 × 1 model grid points.

Project	Observed	Tremblay scheme				Appleman scheme			
		Forecast		Total	Statistics	Forecast		Total	Statistics
		Yes	No			Yes	No		
CASP II	Yes	436	1104	1540	HR = 0.28	380	1160	1540	HR = 0.25
	No	3618	8228	11 846	FAR = 0.31 TSS = -0.02	3538	8305	11 846	FAR = 0.30 TSS = -0.05
CFDE I	Yes	900	1626	2246	HR = 0.40	586	1660	2246	HR = 0.26
	No	1086	3563	4649	FAR = 0.23 TSS = 0.17	929	3720	4649	FAR = 0.20 TSS = 0.06
CFDE III	Yes	1070	1978	3048	HR = 0.35	562	2486	3048	HR = 0.18
	No	1023	3917	3940	FAR = 0.21 TSS = 0.14	626	4314	4940	FAR = 0.13 TSS = 0.06

FAR and lower TSS. During CFDE I and CFDE III the cloud regions sampled were generally stratiform clouds, often associated with air masses ahead or behind frontal regions. These clouds tended to have warmer cloud-top temperatures and were thinner than those observed during CASP II. As a result, measurements of SCW were more frequent in CFDE I and III clouds. Not accounting for depletion by glaciation is likely a weakness in the SCW forecast schemes. Results from the same sensitivity tests used in the cloud verification are listed in Table 6. The HR and FAR values increased with increasing grid size, as was the case in the cloud forecast validation. The TSS values were generally a maximum for the 25 km × 25 km comparisons.

An intercomparison between the results for the Tremblay and Appleman SCW forecasting schemes (Tables 5 and 6) shows that the Appleman scheme produces smaller TSS scores. This demonstrates quantitatively that the Tremblay scheme is more skilled. Based on CASP II simulations of 18 cases, Tremblay et al. (1996) found that the Appleman scheme tended to overforecast SCW events, particularly when the temperatures were colder than -10°C. Conversely, Table 5 shows a slightly higher FAR for the Tremblay scheme, with the maximum difference observed in the CFDE III analysis. This occurs

because most of the observed and forecast SCW events for CFDE III (88% of the SCW forecasts from the Tremblay scheme and 70% of the SCW forecasts from the Appleman scheme) occur in the temperature range between -10° and 0°C. Tremblay et al. (1996) show that their scheme predicts more SCW events than the Appleman scheme at temperatures warmer than -10°C. To confirm this observation, FAR values were calculated for both schemes for the CFDE III dataset for two temperature ranges: -15.5°C < T < -10°C and -10°C < T < 0°C. The FAR for the Tremblay and Appleman schemes were 0.18 and 0.11, respectively, for data between -15.5° and -10°C, and 0.12 and 0.24, respectively, for data between 0° and -10°C. These results are consistent with the conclusions of Tremblay et al. (1996).

The ability of the model to forecast SCW depends significantly on its ability to determine cloud. It is probable that some of the SCW forecast error was directly associated with errors in forecasting cloud. This will be especially true for cases where the model dynamics caused the cloud forecast to be in error. To provide a fairer analysis of the SCW forecasting algorithm, the 1.5-km resolution CFDE I and CFDE III data were analyzed only for points where the model cloud forecast agreed with the aircraft cloud observed (i.e., forecast

TABLE 6. SCW forecast verifications using multiple grid points for model data.

Project	Grid points	Tremblay scheme			Appleman scheme		
		HR	FAR	TSS	HR	FAR	TSS
CASP II	1 × 1	0.28	0.31	-0.02	0.25	0.30	-0.05
	11 × 11	0.33	0.36	-0.03	0.28	0.35	-0.07
	17 × 17	0.35	0.40	-0.05	0.31	0.39	-0.08
	33 × 33	0.39	0.47	-0.08	0.35	0.46	-0.11
	67 × 67	0.52	0.58	-0.06	0.43	0.59	-0.15
CFDE I	1 × 1	0.40	0.23	0.17	0.26	0.20	0.06
	11 × 11	0.46	0.27	0.19	0.34	0.27	0.07
	17 × 17	0.49	0.30	0.20	0.38	0.29	0.09
	33 × 33	0.55	0.38	0.17	0.47	0.36	0.10
	67 × 67	0.66	0.54	0.12	0.59	0.49	0.10
CFDE III	1 × 1	0.35	0.21	0.14	0.18	0.13	0.06
	11 × 11	0.45	0.29	0.16	0.29	0.18	0.11
	17 × 17	0.49	0.32	0.18	0.33	0.21	0.12
	33 × 33	0.58	0.40	0.18	0.41	0.30	0.11
	67 × 67	0.66	0.52	0.14	0.50	0.40	0.10

TABLE 7. SCW forecast verifications for different cloud threshold values (QC_0), for observations made when the cloud forecast was correct.

QC_0 ($g\ m^{-3}$)	HR	FAR	TSS
0.01	0.63	0.22	0.41
0.02	0.67	0.19	0.49
0.03	0.71	0.16	0.55
0.04	0.77	0.14	0.64

and observed, or not forecast and not observed). With these criteria, the accuracy of the SCW forecast increased significantly. Table 7 lists the HR, FAR, and TSS for several cloud threshold values (QC_0). An increased HR resulted in substantially higher TSS values compared to those in Table 5. For example, for a QC_0 of $0.01\ g\ m^{-3}$, at a resolution of 1.5 km, the TSS increased from 0.15 to 0.41. Increasing the QC_0 induces a higher HR and lower FAR, and therefore, a higher TSS value. This is expected because higher QC_0 values would screen weaker LWC signals into the not observed and not forecast category. While these results illustrate that the HR for SCW forecasts is related to the cloud forecast accuracy, there remains room for improvement. The HR will also be related to the model ability to predict vertical velocity and other meteorological parameters. It is not possible to make a direct comparison between the forecast and observed vertical velocity because vertical velocity is not measured from the aircraft.

Figure 7 shows vertical profiles of HR, FAR, and TSS for the Tremblay scheme for cases when the cloud field was predicted correctly. The HR does not change significantly with height except at the highest level, and in general, the HR values are between 0.6 and 0.7. The FAR tends to increase with height, which is likely proportional to decreasing temperature. Overall, the SCW forecast appears highly skilled in regions where the model and aircraft cloud measurements agree.

6. Conclusions

In this paper, in situ aircraft measurements of Ta, Td, V, and TWC, and assessments of cloud and SCW fields, made during 50 flights from three research field programs, have been compared to forecasts made with the GEM HIMAP model. The model and aircraft comparisons were made on a 1.5-km scale, with the aircraft data being averaged from a smaller scale and the model data being numerically interpolated from a larger scale. The comparisons have led to a number of conclusions:

- 1) The rms errors for Ta, Td, and V agreed closely with the results from radiosonde–model validation experiments. This implies that the aircraft–model validation methodology is equally feasible and can be applied to parameters such as TWC, which the radiosondes cannot measure.
- 2) The HR, FAR, and TSS for the Sundqvist cloud scheme were 0.52, 0.30, and 0.22, respectively, when

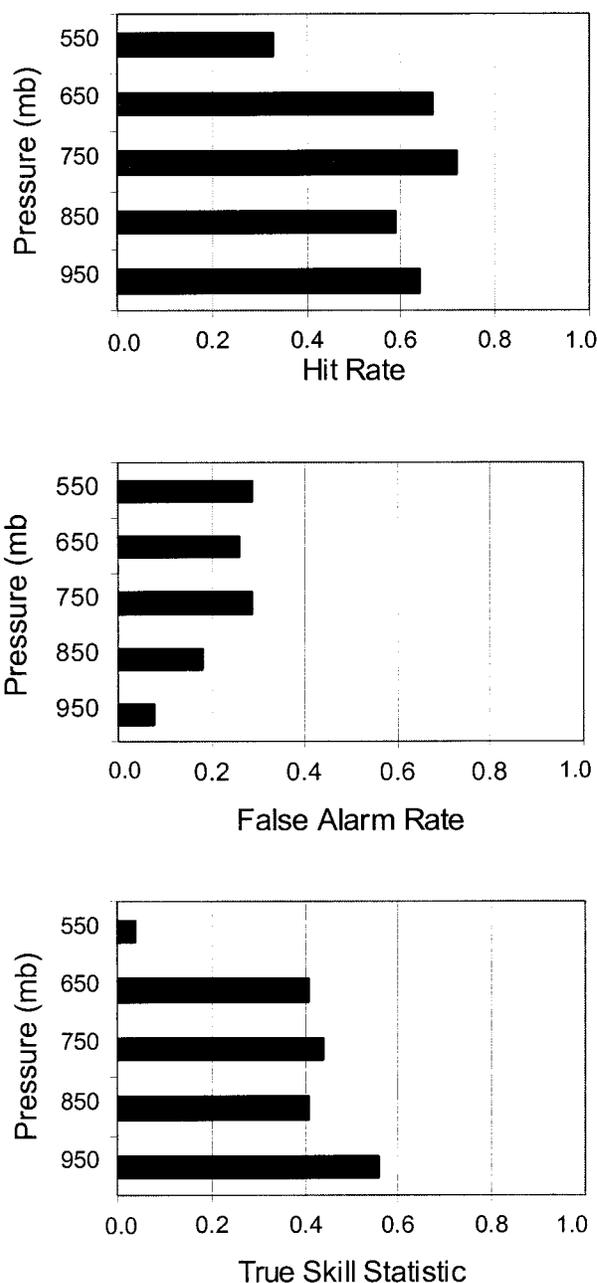


FIG. 7. Vertical distributions of HR, FAR, and TSS for the Tremblay SCW algorithm.

the model data were inferred at a resolution of 1.5 km. The corresponding values increase to 0.76, 0.50, and 0.26, respectively, when a model resolution of 100 km is used. Based on the TSS, the forecast accuracy for low-level cloud is poorer than for middle- and high-level cloud. This is consistent with the results of Yu et al. (1997) who showed that boundary layer clouds were poorly handled in the model.

3) A quantitative comparison of the forecast and observed TWC indicates a poor agreement with the

- model generally underestimating TWC. While forecast errors from model dynamics and errors inherent in the comparison methodology will account for some of the observed scatter, the results suggest that the microphysical parameterizations of cloud TWC in the Sundqvist scheme do not accurately resolve cloud TWC.
- 4) The HR, FAR, and TSS for SCW forecasts based on the Tremblay SCW forecasting scheme, at a resolution of 1.5 km, are 0.37, 0.22, and 0.15, for CFDE I and CFDE III combined. The HR, FAR, and TSS change to 0.66, 0.53, and 0.13, respectively, when a model resolution of 100 km is used. A comparison between the Tremblay and Appleman SCW forecasting schemes demonstrates that the Appleman scheme produces a lower HR, FAR, and TSS.
 - 5) When only cloud forecasts that agree with the aircraft measurements are considered, the accuracy of the Tremblay SCW forecasting algorithm increases significantly, with HR (TSS) increasing from 0.37 (0.15) to 0.63 (0.41). This illustrates that a limiting factor in the SCW forecasts is the model cloud field. A more sophisticated cloud scheme with explicit liquid and ice phases may improve the forecast accuracy.
 - 6) The comparison methodology described here can provide a quantitative intercomparison technique for several SCW or cloud algorithms. Considering that the results demonstrate that improved cloud and SCW algorithms are required, such a comparison would allow identification of strengths and weaknesses in these algorithms, which could lead to suggestions for improving SCW forecast accuracy. Incorporating additional field project data would help address the statistical and geographical limitations of the dataset. Finally the comparison methodology might be expanded to address a wider range of forecasts. For example, rather than segregating the forecasts into yes or no, the SCW forecasts might be divided into bins of different TWC, or categories such as light, moderate, and severe. Such investigations are currently being undertaken.
- Acknowledgments.* The authors are grateful to Richard Moffet and Sylvie Gravel for their support in using the GEM code, and to Marie-France Turcotte for her help in using the SCW forecasting codes. Andre Tremblay is thanked for his advice and careful reviews of the manuscript. This work was supported by the Canadian National Search and Rescue Secretariat.
- REFERENCES
- Appleman, H., 1954: Design of a cloud-phase chart. *Bull. Amer. Meteor. Soc.*, **35**, 223–225.
- CMC, cited 1998: Verification of objective forecasts of weather elements. [Available online at http://www.cmc.ec.gc.ca/~cmsw/cmsw/revue/html/elements/html/liste_an.html.]
- Cober, S. G., G. A. Isaac, and J. W. Strapp, 1995: Aircraft icing measurements in East Coast winter storms. *J. Appl. Meteor.*, **34**, 88–100.
- , —, A. V. Korolev, J. W. Strapp, and D. L. Marcotte, 1999: Measurements of aircraft icing environments which include supercooled large drops. *37th Aerospace Sci. Meeting and Exhibit*, Reno, NV, AIAA 99-0494.
- Côté, J., J. Desmarais, S. Gravel, A. Méthot, A. Patoine, M. Roch, and A. Staniforth, 1998a: The operational CMC-MRB Global Environmental Multiscale (GEM) model. Part II: Results. *Mon. Wea. Rev.*, **126**, 1397–1418.
- , S. Gravel, A. Méthot, A. Patoine, M. Roch, and A. Staniforth, 1998b: The operational CMC-MRB Global Environmental Multiscale (GEM) model. Part I: Design considerations and formulation. *Mon. Wea. Rev.*, **126**, 1373–1395.
- Doswell, C. A., R. Davies-Jones, and D. L. Keller, 1990: On summary measures of skill in rare event forecasting based on contingency tables. *Wea. Forecasting*, **5**, 576–585.
- Flueck, J. A., 1987: A study of some measures of forecast verification. Preprints, *10th Conf. on Probability and Statistics in Atmospheric Sciences*, Edmonton, AB, Canada, Amer. Meteor. Soc., 69–73.
- Forbes, G. S., Y. Hu, B. G. Brown, B. C. Bernstein, and M. K. Politovich, 1993: Examination of conditions in the proximity of pilot reports of icing during STORM-FEST. Preprints, *Fifth Int. Conf. on Aviation Weather Systems*, Vienna, VA, Amer. Meteor. Soc., 282–286.
- Isaac, G. A., S. G. Cober, A. V. Korolev, J. W. Strapp, A. Tremblay, and D. L. Marcotte, 1998: Overview of the Canadian Freezing Drizzle Experiment I, II and III. Preprints, *Cloud Physics Conf.*, Everett, WA, Amer. Meteor. Soc., 447–450.
- , —, —, —, —, and —, 1999: Canadian Freezing Drizzle Experiment. *37th Aerospace Sci. Meeting and Exhibit*, Reno, NV, AIAA 99-0492.
- Knapp, D. I., 1992: Comparison of various icing analysis and forecasting techniques. Verification Rep., Air Force Global Weather Center, 5 pp. [Available from Air Force Global Weather Center, 106 Peacekeeper Dr., Offutt AFB, NE 68113-4039.]
- Korolev, A. V., J. W. Strapp, G. A. Isaac, and A. N. Nevzorov, 1998: The Nevzorov airborne hot-wire LWC-TWC probe: Principle of operation and performance characteristics. *J. Atmos. Oceanic Technol.*, **15**, 1495–1510.
- Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.*, **30**, 291–303.
- Moninger, W. R., and Coauthors, 1991: Shootout-89, a comparative evaluation of knowledge-based systems that forecast severe weather. *Bull. Amer. Meteor. Soc.*, **72**, 1339–1354.
- Panofsky, H. A., and G. W. Brier, 1965: *Some Applications of Statistics to Meteorology*. Pennsylvania State University Press, 224 pp.
- Rasmussen, R., and Coauthors, 1992: Winter Icing and Storms Project (WISP). *Bull. Amer. Meteor. Soc.*, **73**, 951–974.
- Schultz, P., and M. Politovich, 1992: Toward the improvement of aircraft icing forecasts for the continental United States. *Wea. Forecasting*, **7**, 491–500.
- Sundqvist, H., E. Berge, and J. E. Krisjansson, 1989: Condensation and cloud parameterization studies with a mesoscale numerical weather prediction model. *Mon. Wea. Rev.*, **117**, 1641–1657.
- Swets, J. A., 1988: Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293.
- Tremblay, A., A. Glazer, W. Szyrmer, G. A. Isaac, and I. Zawadzki, 1995: Forecasting of supercooled clouds. *Mon. Wea. Rev.*, **123**, 2098–2113.
- , S. G. Cober, A. Glazer, and G. A. Isaac, 1996: An intercomparison of mesoscale forecasts of aircraft icing using SSM/I retrievals. *Wea. Forecasting*, **11**, 66–77.
- Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 467 pp.
- Yu, W., L. Garand, and A. P. Dastoor, 1997: Evaluation of model clouds and radiation at 100 km scale using GOES data. *Tellus*, **49A**, 246–262.