P7.3

CLOUD CEILING HEIGHT ESTIMATION USING GOES-10 AND COAMPS™ DATA

Richard L. Bankert* and Michael Hadjimichael
Naval Research Laboratory, Monterey, CA

1 INTRODUCTION

The U.S. Navy operational meteorologist is often required to assess and forecast cloud ceiling conditions at remote locations where there are no direct observations available. In these situations, generating an immediate diagnosis of cloud ceiling would enhance the support these meteorologists provide to the tactical decision-making process.

Neither numerical model output nor satellite imagery can reliably provide the cloud ceiling height at a specific location. Algorithms such as the one presented in Stoelinga and Warner (1999) (referred to hereafter as SW) are applied to NWP model output in order to extract cloud ceiling fields from numerical model fields. SW is based on empirical and theoretical relationships between hydrometer attributes and light extinction. As noted in that article, SW ceilings were consistently higher than observed. Additionally, satellite data techniques have been developed and applied with limited success. Ellrod (2002) used a surface temperature and infrared cloud top temperature difference to estimate low cloud ceiling heights at night.

In general, the modeling of weather phenomena has been theory-driven: parameters are determined by equations developed from physical laws and subsequently verified by data. However, in some highly complex situations, the physical laws governing these phenomena are either unknown, too complicated to represent, or not fully understood. For example, while a system of physical equations allows for the modeling of temperature and winds, phenomena such as cloud processes are more complex and their relevant parameters must be approximated. In these circumstances, the conceptual modeling can be data-driven. That is, through proper analysis, data relationships representing the physics implicit in the data are empirically discovered.

Supervised machine learning techniques are used to discover patterns in data and to develop associated classification and parameter estimation algorithms. These data mining methods, used in a Knowledge Discovery from Databases (KDD) procedure, are applied to the cloud ceiling height assessment problem. Within the KDD methodology, raw data are collected, processed, and stored in a database. Data mining tools are then applied to the database records to uncover the patterns and relationships that represent physical laws implicit in the data. This research attempts to find relationships in satellite and NWP data that can provide accurate estimates of cloud ceiling height.

COAMPS™ (Coupled Ocean/Atmosphere Mesoscale Prediction System) output, GOES-10 data, and ground observations (METAR) are established within a unique meteorological research environment that allows for the automated collection, processing, and storage of meteorological data records. Parameter values from the disparate data sources are extracted with location and time markers and these values are combined ("fused") to form a single data record with elements that are coincident in space and time. Hourly data records are included in a single database optimized for data mining. In addition to the automated routines to populate the database, this research environment also includes web-based monitoring tools and 2D images of satellite and NWP data that can be accessed in near real-time. The monitoring tools provide supplemental information related to individual NWP model runs, missing data (from any source), and database statistics. This unique KDD environment has been introduced in Bankert et al. (2001).

Parameter values for each data type are collected for 18 METAR (Aviation Routine Weather Report) station locations in California. Cloud ceiling observations are parsed from METAR (mostly automated) reports for these selected stations and stored in the database. They represent ground truth and serve as the dependent variable in the subsequent search for patterns in the data which relate GOES-10 and COAMPS™ variables to locally observed cloud ceiling height.

2 DATA DESCRIPTION

GOES-10 and COAMPS™ each provide a unique type of data. Satellite imagery provides a view "from above" in the form of actual observed radiances in various spectral channels. A numerical model provides a large number of calculated variables at multiple levels in an atmospheric column at any particular model grid point. These data sources provide coincident (temporal and spatial) data that

* *Corresponding author address:* Richard L. Bankert, Naval Research Laboratory, Marine Meteorology Division, 7 Grace Hopper Ave., Monterey, CA 93943; email: bankert@nrlmry.navy.mil

can be explored individually (NWP-only or satellite-only) or in combination. COAMPS™ output parameters and coincident GOES-10 parameters are computed and extracted at 18 METAR observation sites (listed in Table 1). Automated data collection routines allowed for data to be collected hourly over a 2.5 year period (July 12, 2000 - November 29, 2002).

Table 1: The 18 METAR stations (station identifiers in parentheses) in California used in this study. The number of observation records in the database for each location is also indicated.

Location	Records
Napa (KAPC)	26830
Bakersfi eld (KBFL)	24926
Camarillo (KCMA)	26472
Los Angeles (KLAX)	26006
Lompoc (KLPC)	62846
Modesto (KMOD)	25139
Monterey (KMRY)	29011
Paso Robles (KPRB)	25306
San Diego (KSAN)	26219
Santa Barbara (KSBA)	26329
San Luis Obispo (KSBP)	27250
San Francisco KSFO	24802
San Jose (KSJC)	23760
Santa Maria (KSMX)	29597
Salinas (KSNS)	27425
Santa Rosa (KSTS)	29890
Van Nuys (KVNY)	24635
Lancaster (KWJF)	27724

2.1 *COAMPS™ Data*

COAMPS™ is a non-hydrostatic, multiply nested, mesoscale NWP model. It is run over the U.S. West Coast and configured with three horizontal nested grids - 81, 27, and 9 km resolution (Figure 1). There are 33 vertical levels for all grids with the top located at 32.1 km. Grid points are strongly compressed near the surface to resolve the shallow boundary layer. COAMPS™ is run for a 12-hour forecast cycle for this domain configuration at 00 UTC and 12 UTC each day. The Navy Operational Global Atmospheric Prediction System (NOGAPS) provides the time-dependent boundary conditions for the 81 km domain. The model is described by Hodur (1997), Hodur et al. (2002), and Chen (2003). COAMPS™ features a full suite of physical parameterizations, including the Mellor and Yamada (1982) level 2.5 turbulence parameterization, radiation (Harshvardhan et al., 1987), and cloud microphysics (Rutledge and Hobbs, 1983) schemes.

The closest land grid point (within the 9 km domain) to each of the selected METAR stations is determined. COAMPS™ output values at those grid points for each hour are extracted and written to the database. Table 2 is a list of the COAMPS™ parameters utilized for the present study.

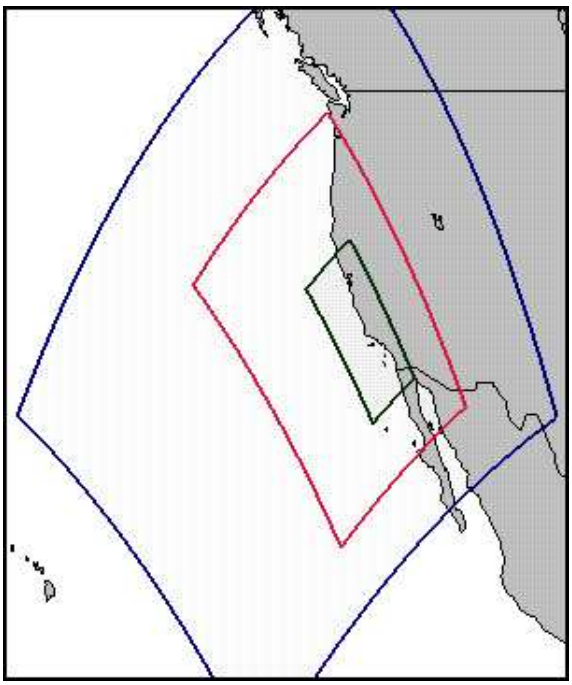

Figure 1: COAMPS™ triply-nested domains for the U.S. West Coast with horizontal grid resolutions of 81, 27, and 9 km for the outermost, middle, and innermost grid, respectively.

COAMPS™ parameters were selected based on a priori assumptions about which parameters might have the most influence on cloud ceiling height. Most of the descriptions in Table 2 are self-explanatory. However, some require additional attention: u_*, t_*, q_* are the friction scale velocity, temperature, and moisture from the Monin-Ohbukhov similarity treatment of the surface layer; the “10 m, sfc temperature (mixing ratio) difference” is the difference in degrees (gm/Kg) between the value at 10 m and the value at the surface; the cloud parameters, “Cloud base height (qc)” and “Cloud top height (qc)”, are determined from examination of the prognostic cloud liquid water and cloud ice mixing ratio field while those followed by “(RH)” are based on relative humidity; LCL is the lifting condensation level and CCL is the convective condensation level; z/L is the height of the surface layer divided by the Monin-Ohbukhov length scale (in COAMPS™, the height of the lowest model grid point is used for the depth of the surface layer as is the common practice in mesoscale models); and the cloud/no cloud determination is based on the existence of cloud liquid water or cloud ice in the atmospheric column above the point of interest.

While COAMPS™ cannot forecast cloud ceiling height with sufficient reliability and accuracy to be of operational

use, three dervied COAMPS™ parameters represent the cloud ceiling height at each grid point. One of these ceiling height estimations, “Cloud base height (qc)”, is based on cloud water and ice. A search is made in a vertical column for the lowest altitude at which cloud liquid water mixing ratio or the cloud ice mixing ratio exceeds a threshold of 1.0e-06. A second COAMPS™ ceiling height parameter is cloud base height (RH). The method for computing this height is similar to cloud base height (qc) except that a search is made for the lowest height level at which the relative humidity is greater than 95%. Ceiling height (SW) first uses the mixing ratios for cloud liquid water, cloud ice, snow, and rain water to compute concentrations for each species. The extinction coefficients are then computed and integrated upward to the lowest altitude at which a light beam from the surface decreases to 0.02 times the original intensity.

2.2 *GOES-10 Data*

Hourly GOES-10 pixel data are extracted and written to the database. This data consists of all sensor channel data at a given pixel whose center (over land) is closest to the location of each of the METAR stations. The visible channel value is corrected for the solar zenith angle. A cloud optical depth algorithm (Wetzel and Stowe, 1999) is applied to the GOES-10 data (daytime only) and a low cloud product (Lee et al., 1997) is derived by computing the difference between the shortwave and longwave infrared (IR) channels. In addition, the difference in the two longwave IR channels is computed. Tables 3 and 4 summarize this information and include resolution and coverage information.

2.3 *METAR Data*

METAR reports were collected in near-real time each hour from the Fleet Numerical Meteorological and Oceanographic Center (FNMOC) data server. Observed cloud ceiling height is one of the sensible weather elements parsed from the hourly METAR reports for the 18 selected stations. The map in Figure 2 is marked with the locations and names of those selected stations. These METAR stations were chosen due to their coastal nature (for most stations), the availability of satellite data over their location, and the reliability and robustness of the METAR reports. The ceiling height values represent the observer ground truth and serve as the dependent variable in the regression relating GOES-10 and/or COAMPS™ variables to local cloud ceiling height. The observed cloud ceiling height is defined as the lowest level that has at least broken sky conditions (equal to or greater than 6/8 cloud coverage).

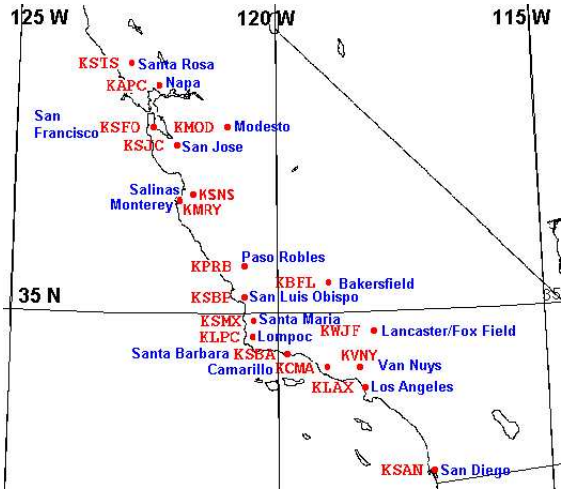

Figure 2: Inner (9km) grid METAR station names and locations.

3 DATA MINING PROCEDURE

To collect, visualize, interpret, and exploit the vast amount of digital data available in the environmental sciences, researchers have turned to Artificial Intelligence (AI) methods, and specifically, data mining (Hand et al., 2001). Data mining is a discipline born of machine learning and statistics, enhanced by large database concerns, pattern recognition, knowledge representation, and other areas of computer science and AI. In contrast to standard statistical approaches, data mining methods in the KDD tool chest typically relax requirements of sample size and pre-specified model hypotheses. They are driven by the data and do not require the correct preselection of hypothesis. Statistical methods generally hypothesize the form of the model, and then use data to confirm the hypothesis. Furthermore, data mining methods are designed to be able to handle larger data sets (millions of records, and hundreds of variables). As a tool for scientific data analysis, data mining utilizes induction to determine empirical models from the observed data. This is in contrast to traditional methods of analysis, where a hypothesis is made based on understanding of physical laws, and data is used to confirm or refute the hypothesis.

In this study, the principle analysis tools for data mining are the supervised inductive learning tools C5.0 and Cubist (Quinlan, 1993; Rulequest Research, 1997-2004). These tools were selected because of their ease of use and recognized robustness.

C5.0 is a data mining algorithm used for producing classification models in the form of decision trees or *if-then* rules. The software is designed to explore hundreds of thousands of database records with hundreds of numeric fields. As these classification models are expressed as de-

Table 2: COAMPS™ variables used in the data mining process.

1000mb, 850mb thickness	Ground temperature
10m dewpoint	Ground wetness
10m latent heat flux	LCL
10m potential temperature	Max TKE in PBL
10m relative humidity	Max mixing ratio in PBL
10m sensible heat flux	Max vert. velocity in PBL
10m temperature	Net radiation
10m u-wind	PBL depth
10m v-wind	Precipitable water
10m, 1500m temp diff	Sea level pressure
10m, sfc mixing ratio diff	Surface albedo
10m, sfc temperature diff	Surface mixing ratio
Bulk Richardson number	Surface roughness
CCL	Surface wind stress
Ceiling height (SW)	Topography height
Cloud base height (RH)	Total downward radiation
Cloud base height (qc)	Total heat flux
Cloud coverage	z/L
Cloud top (qc) temperature	q_*
Cloud top height (qc)	t_*
Cloud/No Cloud	u_*

Table 3: GOES-10 Sensor channel information as captured for the database.

Channels	*Central Wavelength*	*Resolution*	*Coverage*
1. Visible	.65 μm	1 km	30 Minutes
2. Near IR	3.9 μm	4 km	
3. IR (Water Vapor)	6.7 μm	8 km	
4. IR (Thermal)	11.0 μm	4 km	
5. IR (Thermal)	12.0 μm	4 km	

cision trees or rules, they are easier to interpret than other "black-box" data mining tools such as neural networks.

In this work, C5.0 is used to generate two types of classifiers. The first classifier classifies Event records into *ceiling* and *no-ceiling* categories. The second classifies *ceiling* records into *high-ceiling* and *low-ceiling* categories.

The Cubist algorithm produces rule-based predictive models for numerical prediction (also known as regression). Each model is expressed as a set of rules. Each rule applies to only a small part of the input space, and has a set of preliminary conditions and an associated local multivariate linear model. If a rule's conditions are satisfied, the associated model is used to calculate the predicted value. This approach works well in high-dimension problems, such as the one addressed in this work, as only a small number of variables may be required for a particular rule and model. As a result, the rule set is more easily interpreted than a standard regression equation on all input variables, or a neural network.

3.1 *Experiment Methodology*

Through a KDD process, a 3-step method was developed for generating cloud ceiling classifiers and ceiling height estimators:

Step 1 : Create a Cloud Ceiling / No Cloud Ceiling classifier (using C5.0 with all training data).

Step 2 : Create a Low Cloud Ceiling / High Cloud Ceiling classifier (using C5.0 with training data consisting only of cases where a cloud ceiling is present).

Step 3 : Create a Low Cloud Ceiling height estimator (using Cubist with training data consisting only of cases where a low cloud ceiling is present).

Table 4: GOES-10 product information.

Product	*Coverage*
Sun angle corrected visible channel	30 Minutes
Cloud Optical Depth	30 Minutes
Low Cloud Product	30 Minutes
Longwave IR Difference	30 Minutes

The resulting system is executed analogously. If the result of Step 1 classification for a given data point is "Cloud Ceiling," then the Step 2 classifier is executed for that point. The Step 3 estimator is executed if the result of the Step 2 classification is "Low Cloud Ceiling."

The "No Cloud Ceiling" class in the first step also includes data records for which the METAR ceiling is above 12,000 ft (3657.6 m). Most METAR reporting stations are automated and observation instruments at such stations cannot detect clouds above that altitude. Therefore, if the lowest observed ceiling is above this limit, it is indistinguishable from a "No Cloud Ceiling" condition. As a result, all observed ceilings greater than 3657.6 m are classified as "No Cloud Ceiling." For Step 2, the threshold to separate low and high ceilings is 1000 m.

The basic classifier/estimator algorithm development was performed using the following procedures:

- Export a set of Event records data from the database.
- Randomly split data into equal-sized training and testing data sets
- Perform C5.0 data mining for Step 1 on training data, and test resultant algorithm on testing data.
- Perform C5.0 data mining for Step 2 on "Ceiling" cases only in the training data, and test the resultant algorithm on "Ceiling" cases from the testing data.
- Perform Cubist data mining for Step 3 on "Low Cloud Ceiling" cases only in the training data, and test the resulting algorithm on the "Low Cloud Ceiling" cases in the testing data.
- When satisfied with results, output an algorithm trained on all available data for incorporation into the final, production algorithm.

All learning experiments were based on three different sets of variables:

1. COAMPS™ variables only.
2. GOES-10 variables only.
3. Fused COAMPS™ and GOES-10 variables.

In addition to splitting the data into training and testing sets, the data were further separated into day and night as determined by the solar zenith angle (where an angle of less than or equal to 80° is considered daytime). As mentioned previously, complete hourly records for each of the 18 METAR stations were collected over a 2.5 year period. Complete records used for performance evaluation included COAMPS™ output, GOES-10 data, and the METAR cloud ceiling height. There are 263,483 complete records for the California stations.

4 RESULTS

For each of the three steps defined in Section 3.1, four algorithms are compared in terms of bias, accuracy, and skill on the daytime data for all METAR stations. The four algorithms are

1. KDD-produced algorithm using GOES-10 data.
2. KDD-produced algorithm using COAMPS™ data.
3. KDD-produced algorithm using both GOES-10 and COAMPS™ data.
4. SW translation algorithm applied to COAMPS™ data.

For Step 1 (ceiling/no-ceiling classification), there are 51,611 randomly-selected training records and 51,690 randomly-selected testing records. The training records are used to create the algorithm. Table 5 is a listing of the performance statistics, on the testing set, for Step 1.

In Table 5, bias is defined as the ratio of "predicted" ceilings over observed ceilings. All four algorithms tested produced a bias value less than 1.0 (Table 5), indicating an underprediction of ceiling events or a bias toward no ceiling classification. This bias is particularly strong for SW.

To measure the accuracy of the algorithms in determining cloud ceiling events, the percent correct (% correct), probability of detection (POD), false alarm ratio (FAR), and critical success index (CSI) (Marzban, 1998) are computed and presented in Table 5. These measures are described as follows, with "event" defined as ceiling for this step:

% correct is the total percentage correct for both events and non-events.

POD (Probability of Detection) is the fraction of observed events that were correctly predicted to exist. Ignores false alarms.

Table 5: Ceiling/No Ceiling classification performance statistics for the four algorithms used on the California daytime testing data set. POD: Probability of detection, FAR: False alarm ratio, CSI: Critical success index, ETS: Equitable threat score, TSS: True skill score.

Algorithm	Bias	% correct	POD	FAR	CSI	ETS	TSS
KDD NWP+SAT	.91	93	.80	.12	.72	.66	.77
KDD SAT	.91	92	.78	.14	.69	.62	.74
KDD NWP	.83	87	.61	.26	.51	.42	.56
SW	.34	81	.23	.32	.21	.15	.20

FAR (False Alarm Ratio) is the fraction of predicted events that are non-events. Ignores missed events.

CSI (Critical Success Index), also known as the threat score, is the ratio of correctly predicted events (hits) with the total number of hits, misses, and false alarms. Does not distinguish source of forecast error.

By any of these accuracy measurements (Table 5), the combination of GOES-10 and COAMPS™ in the KDD cloud ceiling algorithm produced the best results. Satellite-only (GOES-10) algorithm scores are only slightly lower. All three KDD algorithms are much more accurate than SW.

The skill scores computed here include:

ETS (Equitable Threat Score) is a measure of skill that uses chance as the benchmark. Accounts for climatological event frequencies. Range of values is -.333 to +1.0 (0 is no skill).

TSS (True Skill Score - (Hanssen and Kuipers, 1965)) examines the ability of the algorithm to separate events from non-events (accuracy of events + accuracy of non-events - 1.0), with scores ranging from -1.0 to +1.0. Does not depend on data distribution. The benchmark for this score is the “naive” prediction, e.g, event always (or never) predicted produces a TSS of 0.0 (no skill).

Based on these skill scores (Table 5), the KDD algorithm incorporating both GOES-10 and COAMPS™ data demonstrated the most skill. All three KDD cloud ceiling algorithms have much higher skill scores than SW.

For Step 2, low/high cloud ceiling classifications (with low cloud ceiling as the event being analyzed), there are 11,279 randomly-selected training samples and 11,199 randomly-selected testing samples. Similar to Step 1, bias, accuracy, and skill are computed and presented in Table 6.

The bias values of the four algorithms for this classification step indicate a very slight bias toward low ceilings in the KDD algorithms and a bias toward high ceilings for SW.

Accuracy measurements shown in Table 6 are similar for the three KDD-produced algorithms with the fused-data algorithm having slightly better results. While SW has comparable FAR to the KDD algorithms, the other three accuracy scores are much lower. These results indicate the KDD cloud ceiling algorithms minimize both misses and false alarms, but SW frequently misses low ceiling events.

Similar conclusions can be drawn from the skill scores (Table 6) at this step. Compared to Step 1 (Table 5), the skill level in Step 2 (Table 6) is much lower for all algorithms except the KDD algorithm using only COAMPS™ data. This algorithm actually has higher scores for the low/high ceiling classification. Information from the atmospheric column as seen in the COAMPS™ data is helpful in distinguishing low ceilings from high. Contrast that result with the KDD cloud ceiling algorithm using only GOES-10 data. Since the satellite data features are representing the atmosphere from an above-cloud perspective only, the skill in detecting the existence of a ceiling is expected to be higher than determining whether the ceiling is high or low.

For the third step of each algorithm, the heights of the low ceiling (less than 1000 m) cases (training - 8429 samples; testing - 8389 samples) are estimated. Performance measures are presented in Table 7 for this step. The bias computed for this step is equivalent to the average error of the testing set. The KDD-produced algorithms have very small negative bias and SW has a more substantial positive bias (Table 7).

The three accuracy measures in Table 7 were computed as follows:

CC (Correlation Coefficient) is a measure of the relationship of the algorithm output with observation. Values range from -1.0 (perfect negative correlation) to +1.0 (perfect positive correlation). A value of 0.0 is no correlation.

MAE (Mean Absolute Error) is the average difference between the algorithm output and the observation for

Table 6: Low Ceiling/High Ceiling classification performance statistics for the four algorithms used on the California daytime testing data set.

Algorithm	Bias	% correct	POD	FAR	CSI	ETS	TSS
KDD NWP+SAT	1.01	87	.92	.09	.84	.47	.63
KDD SAT	1.05	83	.91	.13	.80	.36	.50
KDD NWP	1.01	86	.91	.10	.83	.45	.62
SW	.20	36	.17	.13	.17	.03	.10

the testing set.

RMSE (Root Mean Square Error) is

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(e_i - o_i)^2} \qquad (1)$$

where N is the number of testing samples, e_i is the estimated ceiling height for each testing sample, and o_i is the observed ceiling height for each sample.

RMSE is affected more by larger errors than MAE.

The four CC values (Table 7) are similar, ranging from .57 for SW to .76 for the KDD algorithm using both data sets (GOES-10 and COAMPS™). However, the MAE and RMSE are much lower for the KDD algorithms (Table 7) with the combined data and COAMPS™-only data producing similar MAE and RMSE values. To measure the skill of the KDD cloud ceiling algorithms relative to SW, the following equation was used:

$$\text{Skill} = 1 - \frac{\text{MAE}_{\text{KDD}}}{\text{MAE}_{\text{SW}}} \qquad (2)$$

This skill score provides a single value (with a maximum score of 1.0) of the KDD-produced algorithms' skill level relative to SW (Table 7). As was the case with the first two steps, the KDD algorithm using both GOES-10 and COAMPS™ data demonstrated the most skill. Overall, the KDD "fused"-data algorithm had only slightly better testing results than the KDD GOES-10-only algorithm in Step 1 and the KDD COAMPS™-only algorithm in Steps 2 and 3. However, when all 3 steps are examined as a total cloud ceiling estimation algorithm, the KDD fused-data algorithm provides the most accurate and highest skill in ceiling diagnosis. Table 8 provides a quick-look summary of the skill scores (Steps 1 and 2) and correlation (Step 3) for all four algorithms. Further testing results and algorithm analysis can be found in Bankert et al. (in press).

5 GOES-10 CLOUD CEILING ALGORITHM EXAMPLE

The GOES-10-only KDD cloud ceiling algorithm is being applied hourly for the southern California coast. The images in Figure 3 are the GOES-10 visible channel images over Southern California for 8 June 2004 at 1500 UTC, 2100 UTC, and 2200 UTC, respectively from top to bottom. All of the channel and derived (cloud optical depth, etc) values at a given pixel (4 km resolution) are entered into the KDD cloud ceiling height estimation algorithm and the corresponding ceiling height output images are displayed in Figure 4. Pixels with no ceiling are black. High ceilings (greater than 1000 m or 3280 ft) are marked as white. Low cloud ceiling pixels are colored using the scale provided.

These cloud ceiling height images provide a good example of the ceiling height rising throughout a given day or the ceiling dissipating entirely from early morning to afternoon. KVNY (Van Nuys), location marked on the 1500 UTC image of Figure 3, reported a cloud ceiling of 1800 ft at 1451 UTC, 3900 ft at 2051 UTC, 4100 ft at 2151 UTC. KLAX (Los Angeles), location marked on the 1500 UTC image of Figure 3, reported a ceiling height of 2100 ft at 1450 UTC and no ceiling at 2050 UTC and 2150 UTC. The 1450 UTC 1300 ft cloud ceiling height at KPOC (LaVerne), location marked on the 2100 UTC image of Figure 3, was raised to 2900 ft and 3000 ft at 2047 UTC and 2147 UTC, respectively. A ceiling height of 1500 ft was reported by KOKB (Oceanside), location marked on the 2100 UTC image of Figure 3, at 1456 UTC, followed by a 4600 ft ceiling at 2056 UTC and no ceiling at 2156 UTC. KSAN (San Diego International Airport), location marked on the 2200 UTC image of Figure 3, reported a cloud ceiling height of 1500 ft at 1451 UTC, 3500 ft at 2051 UTC, and no ceiling at 2151Z. By subjective visual inspection of the KDD algorithm output relative to these station locations, this cloud ceiling height estimation algorithm appears to give a fairly accurate diagnosis of the cloud ceiling situation.

This case study demonstrates the potential usefulness of any point-location validation efforts, but there are pitfalls. For example, KSNA (Santa Ana), location marked on the 2100 UTC image of Figure 3, reported a ceiling height of 1900 ft at 1453 UTC. At 2053 UTC there was no ceiling reported. However, a 2400 ft ceiling was reported at 2153 UTC which was closely followed by a report of no ceiling at 2208 UTC. This type of changing cloud ceiling environment at a specific location makes it difficult for validation of the algorithm as it takes a snapshot at a specific

Table 7: Low Ceiling Height Estimation performance statistics for the four algorithms used on the California daytime testing data set. CC: Correlation coefficient, MAE: Mean absolute error, RMSE: Root mean square error.

Algorithm	Bias (m)	CC	MAE (m)	RMSE (m)	Skill
KDD NWP+SAT	-1.17	.76	120.6	168.0	.75
KDD SAT	-2.19	.64	149.5	189.3	.69
KDD NWP	-2.09	.75	124.2	162.2	.74
SW	50.6	.57	478.5	714.6	—

Table 8: Performance summary for each of the four algorithms at each step. TSS = True Skill Score; CC = Correlation Coefficient.

	KDD SAT	KDD NWP	KDD NWP+SAT	SW
Step 1 (TSS)	.74	.56	.77	.20
Step 2 (TSS)	.50	.62	.63	.10
Step 3 (CC)	.64	.75	.76	.57

time with a 4 km horizontal resolution. Similarly at KWYF (San Diego), location marked on the 2200 UTC image of Figure 3, a 2000 ft ceiling was reported at 2053 UTC, but by 2103 UTC it had risen to 3100 ft. Finally, KLPC (Lompoc), location marked on the 1500 UTC image of Figure 3, reported a 900 ft cloud ceiling at 1515 UTC, no ceiling at 2055 UTC, and a 100 ft ceiling at 2155 UTC. The GOES-10-only KDD algorithm appears to be getting some signal in this area, but the cloud area size and timing aspects may be beyond the limitations of this type of algorithm.

6 CONCLUSION

The results from analysis of the California daytime data set demonstrate the potential and viability of using KDD to develop algorithms from NWP and/or satellite data for estimating cloud ceiling conditions. Taking advantage of the unique characteristics of each data type, a KDD-produced algorithm that applies both COAMPS™ and GOES-10 data performed the best over the entire 3-step cloud ceiling estimation system. All three KDD-produced algorithms performed significantly better than the currently operational SW algorithm.

The initial expectation that a combination of satellite and COAMPS™ data in a KDD-produced algorithm would produce the highest skill scores was met. It is also worth noting that the skill level of the KDD-produced algorithm using satellite data alone to diagnose quantitative cloud ceiling height is remarkably high. Of course, multi-layered cloud situations would present a problem if satellite data was the only data source available.

The opportunities for future research using the current database include, but are not limited to, data mining for a cloud ceiling algorithm using data from all three regions studied (Adriatic and Korea in addition to California) to determine if and how a generalized (not region specific) algorithm can be developed, determining a method to incorporate polar-orbiting satellite data, developing and examining satellite and COAMPS™ combined forecast (i.e, not diagnostic) algorithms, and data mining for other weather elements, including visibility.

ACKNOWLEDGMENTS

The support of the sponsor, the Office of Naval Research, under Program Element 0602435N is gratefully acknowledged. The assistance of Jeff Hawkins, Joe Turk for the establishment of useful satellite data at NRL, John Cook, Sue Chen, Tracy Haack, and other COAMPS™ researchers (all with NRL) is very much appreciated. Melanie Wetzel of the Desert Research Institute is acknowledged for her contribution in the area of cloud optical depth estimation. Thanks also to FNMOC for assistance on acquisition of METAR reports.

REFERENCES

Bankert, R.L., M. Hadjimichael, A.P. Kuciauskas, W.T. Thompson, and K. Richardson, in press: Remote cloud ceiling assessment using data mining methods. *Journal of Applied Meteorology*.

Bankert, R.L., M. Hadjimichael, P.M. Tag, A.P. Kuciauskas, W.T. Thompson, and K.L. Richardson: 2001, Remote weather assessment using fused data and knowledge discovery from databases. *Proceedings, 18th Conf. on*

Weather Analysis and Forecasting, American Meteorological Society, Ft. Lauderdale, FL, 432–435.

Chen, S., 2003: COAMPS™ version 3 model description: General theory and equations. Technical Report NRL/PU/7500-04-448, Naval Research Laboratory, 145 pp.

Ellrod, G.P., 2002: Estimation of low cloud base height at night from satellite infrared and surface temperature data. *Nat. Wea. Dig.*, **26**, 39–44.

Hand, D., H. Mannila, and P. Smyth, 2001: *Principles of Data Mining*. MIT Press.

Hanssen, A.W. and W.J.A. Kuipers, 1965: On the relationship between the frequency of rain and various meteorological parameters. *KNMI Meded. Verhand.*, **81**, 2–15.

Harshvardhan, R., R. Davies, D. Randall, and T. Corsetti, 1987: A fast radiation parameterization for atmospheric circulation models. *J. Geophys. Res.*, **92**, 1009–1015.

Hodur, R.M., 1997: The Naval Research Laboratory's Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS™). *Mon. Wea. Rev.*, **125**, 1414–1430.

Hodur, R.M., J. Pullen, J. Cummings, X. Hong, J.D. Doyle, P. Martin, and M.A. Rennick, 2002: The Coupled Ocean/Atmosphere Mesoscale Prediction System COAMPS™. *Oceanography*, **15**, 88–89.

Lee, T.F., F.J. Turk, and K. Richardson, 1997: Stratus and fog products using GOES-8-9 3.9 micron data. *Wea. and Forec.*, **12**, 664–677.

Marzban, C., 1998: Scalar measures of performance in rare-event situations. *Wea. and Forec.*, **13**, 753–763.

Mellor, G.L. and T. Yamada, 1982: Development of a turbulence closure for geophysical problems. *Rev. Geophys. Space Phys.*, **20**, 851–875.

Quinlan, J.R., 1993: *C4.5: Programs for machine learning*. Morgan Kaufmann Pub., San Mateo, 302 pp.

Rulequest Research, 1997-2004: Cubist. http://www.rulequest.com.

Rutledge, S.A. and P.V. Hobbs, 1983: The mesoscale and microscale structure and organization of clouds and precipitation in mid-latitude cyclones. VIII: A model for the "seeder-feeder" process in warm-frontal rainbands. *J. Atmos. Sci.*, **40**, 1185–1206.

Stoelinga, M.T. and T.T. Warner, 1999: Nonhydrostatic mesobeta-scale model simulations of cloud ceiling and visibility for an east coast winter precipitation event. *J. Appl. Meteor.*, **38**, 385–404.

Wetzel, M.A. and L.L. Stowe, 1999: Satellite-observed patterns in stratus microphysics, aerosol optical thickness, and shortwave radiative forcing. *J. Geophys. Res.*, **104**, 31287–31299.

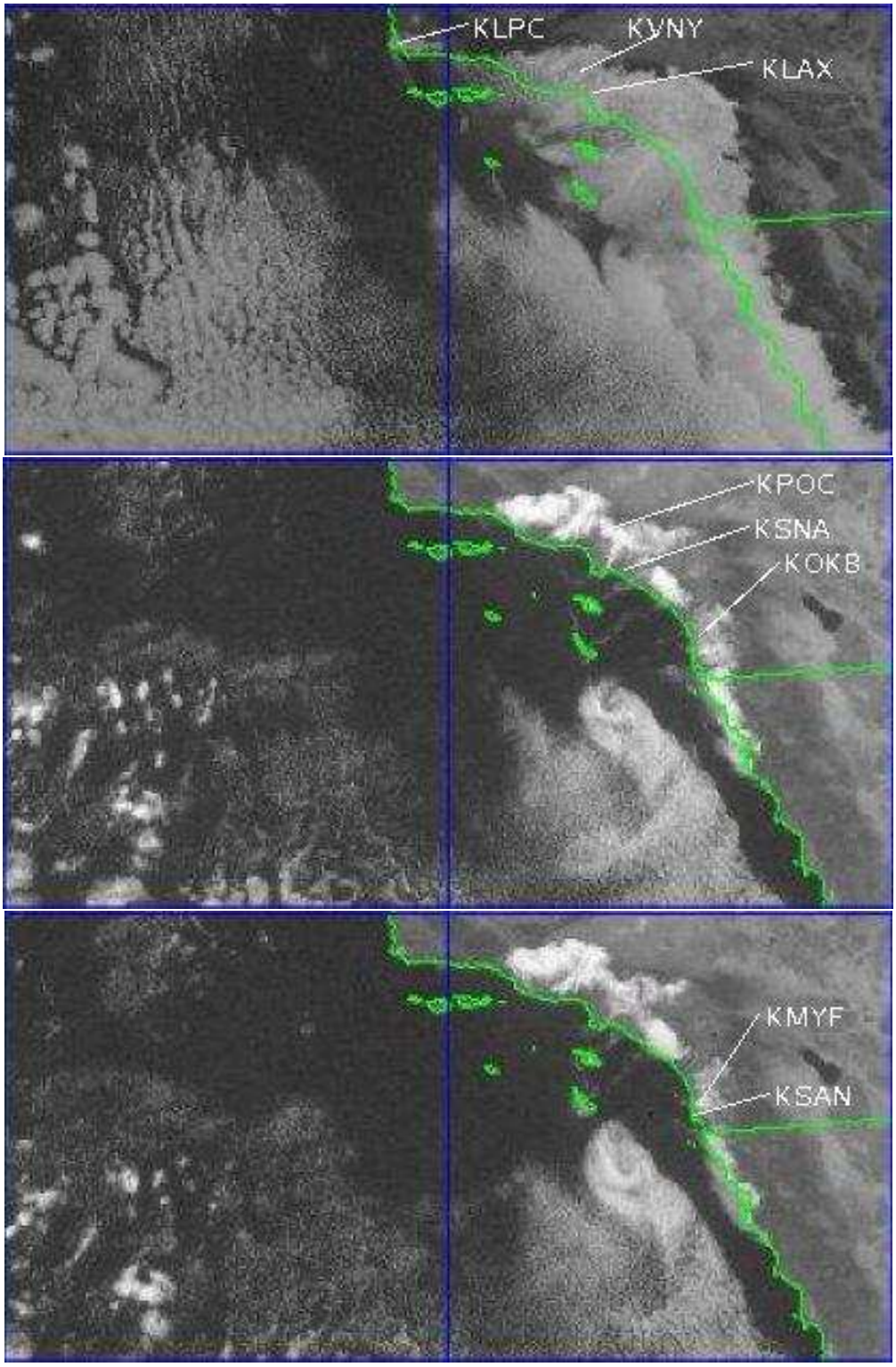

Figure 3: GOES-10 visible images over Southern California and adjacent waters on 8 June 2004 at 1500 UTC (top), 2100 UTC (middle), and 2200 UTC (bottom).

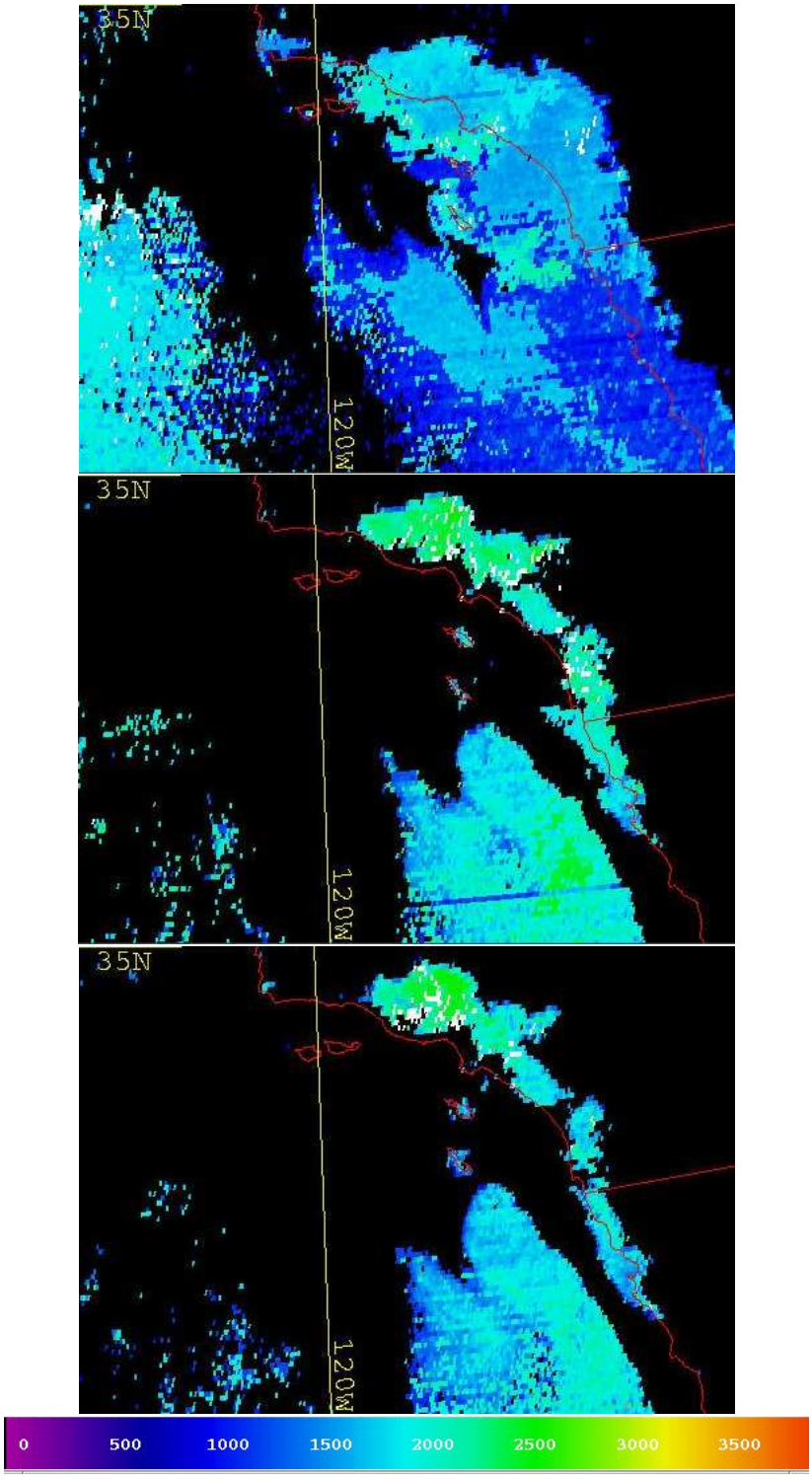

Figure 4: Cloud ceiling heights (ft) images (8 June 2004) as derived from GOES-10 data using KDD-developed algorithm. Top: 1500 UTC; Middle: 2100 UTC; Bottom: 2200 UTC