4.7 APPLICATION OF KNOWLEDGE DISCOVERY FROM DATABASES TO REMOTE WEATHER ASSESSMENT

Michael Hadjimichael*, Richard L. Bankert, Arunas P. Kuciauskas, Kim L. Richardson, Gerard N. Vogel
Naval Research Laboratory, Monterey, California

1. INTRODUCTION

U.S. Navy weather observing and forecasting operations would be greatly assisted with the immediate assessment of remote meteorological parameters when ground observations are not available. To this end, numerical weather prediction data and satellite data from various sensors and platforms are being used to develop automated algorithms to assist in operational weather assessment and forecasting. Supervised machine learning techniques are used to discover patterns in the 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 cloud ceiling height, rain rate, and rain accumulation estimation at remote locations using appropriate geostationary and polar orbiting satellite data in conjunction with Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS) data. Data mining methods have determined an algorithm to diagnose these sensible weather elements more accurately than numerical weather prediction or satellite methods alone. Further detail about the initial design of the study, data, methods, and comparisons to other methods can be found in Hadjimichael el al (1998). Methodology overview and results from the data mining work are presented here.

1.1 Knowledge Discovery from Databases

Data mining is a general term referring to a set of methods for extracting patterns from data. In general, it may apply to both traditional statistical methods, and artificial intelligence machine learning algorithms. Knowledge Discovery from Databases refers to a procedure to using data mining algorithms in a process of studying data to discover useful information (Fayyad et al, 1996a; Weiss and Indurkhya, 1998).

KDD has been successfully applied to many scientific problem, including astronomy, geophysics, and meteorology. Some of the commonly used data mining methods include inductive machine learning, regression, clustering, summarization, generalization, and dependency modeling.

* *Corresponding author address:* M. Hadjimichael, Naval Research Laboratory, Marine Meteorology Division, 7 Grace Hopper Avenue, Monterey, CA 93943-5502; email: hadjimic@nrlmry.navy.mil.

2. DATA SOURCES

In order to discover the relationships between a variety of physical variables, both calculated and measured, a database must be created from a "fusion" of data from various sources.

A unique meteorological research tool consisting of a database of COAMPS output, satellite data, climatology, and ground truth observations (METAR) has being created for use in data mining. COAMPS output parameters, coincident satellite parameters (including both geostationary and polar-orbiting data) and climatological information are extracted/computed at 45 METAR observation sites. Automated data collection routines have been written and data has been collected hourly since July, 2000. Data mining techniques have been applied to study cloud ceiling height and rain accumulation diagnosis.

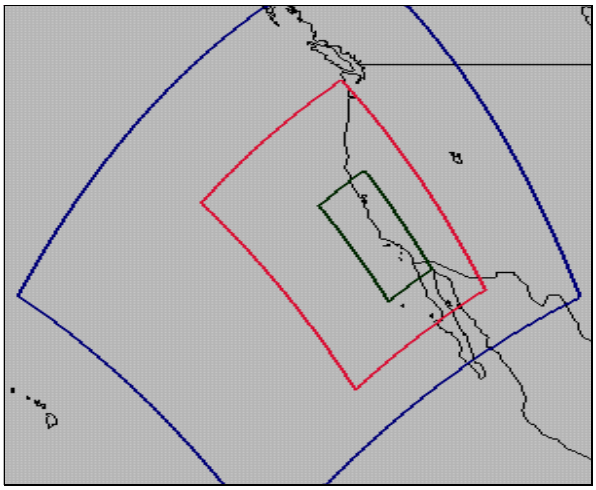

Figure 1. Focus area over U.S west coast. All locations are within the inner grid.

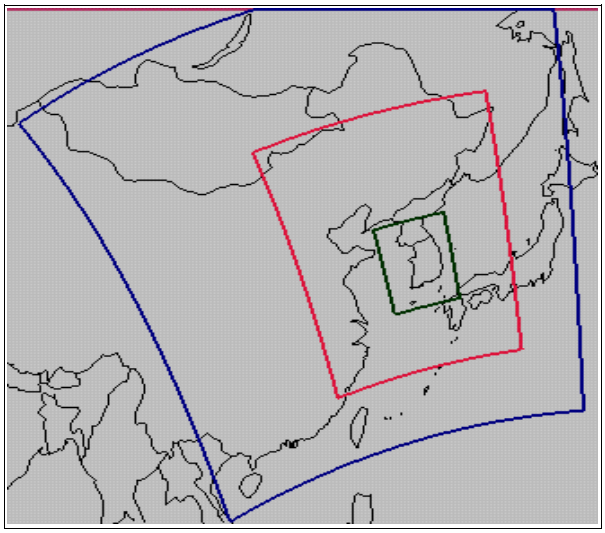

Figure 2. Focus area over Korean peninsula.

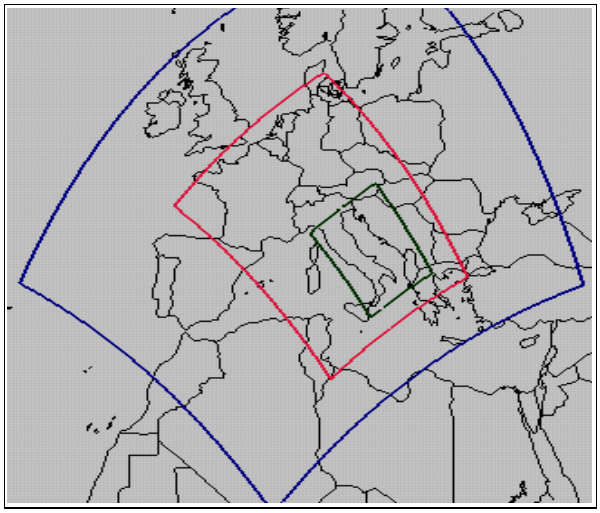

Figure 3. Focus area over Adriatic Sea.

2.1 Numerical Model Data

COAMPS is the numerical weather prediction model used to generate output values of selected relevant parameters. The model is run in three geographic regions, U.S. west coast, Adriatic Sea, and Korean peninsula, and configured with three nested grids (Figure 1). There are 33 grid levels in the vertical. COAMPS is run for a 12-hour forecast cycle for each of these domain configurations at 00 GMT and 12 GMT each day. Interested readers can find additional COAMPS details in Hodur (1997).

The closest land grid point (within each of the 9 km domains) to the 45 (18 West Coast, 14 Adriatic, 13 Korea) METAR stations is determined and COAMPS output values at those grid points for each hour are extracted and written to the database. Table 1 is a list of those COAMPS parameters.

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

Table 1. COAMPS parameters extracted for each of the 45 locations (over 3 focus areas).

In addition to extracting values for the database, COAMPS output can be viewed in static or animated 2D form over the appropriate domains for further interpretation and analysis.

2.2 Satellite Data

Data from three geostationary satellites, GOES-10, European Meteosat-7, and the Japanese GMS-5 are extracted and added to the database. This data will consist of all channel data at a given pixel whose center is closest to the latitude/longitude of each of the METAR stations. All visible channel data is corrected for the solar zenith angle. In addition to the channel data, a cloud optical depth algorithm (Wetzel et al., 1999) is applied and a GOES-only low cloud product (Lee et al., 1997) is derived, with their respective values extracted and stored.

NOAA polar-orbiting Advanced Very High Resolution Radiometer (AVHRR - local area coverage (LAC) and global area coverage (GAC)) data and Defense Meteorological Satellite

Temperature	Wind direction
Vapor pressure	Wind speed
Dewpoint temperature	Wind gust
Altimeter	Weather
Visibility	Total cloud coverage
Ceiling/No ceiling	Lowest cloud coverage
Ceiling height	Cloud coverage fraction and height at all reported levels

Table 2. METAR parameters.

Program (DMSP) Special Sensor Microwave Imager (SSM/I) polar-orbiting data are also extracted and stored in the database. The AVHRR LAC and GOES data records include a derived cloud type classification in addition to the channel data. In addition to the various microwave channel va;lues, environmental data records (EDRs) are computed from the SSM/I channel data These parameters include rain rate, cloud liquid water, and precipitable water.

Using appropriate COAMPS and satellite data, a cloud top height value is derived for all sensors except the SSM/I. Using satellite-based algorithms (Turk et al, 2001) rain rate and accumulation values are copmuted from the geostationary satellite data.

Similar to the COAMPS output, satellite imagery can be viewed in static or animated 2D form. In addition to this visualization, monitoring tools have been developed to allow for a quick view of model and satellite retrieval performance.

2.3 Ground Truth Data

All collected METAR reports for the 45 selected stations are parsed, with sensible weather elements stored in the database. These weather parameters represent the ground truth and are the dependent variables in the subsequent search for patterns which relate satellite and model variables to locally observable parameters. Table 2 is a list of METAR elements. Of the variables in Table 2, Ceiling height is the first parameter examined in the data mining portion of the KDD process. Rain rate and rain accumulation analysis will follow.

3. DATABASE DEVELOPMENT

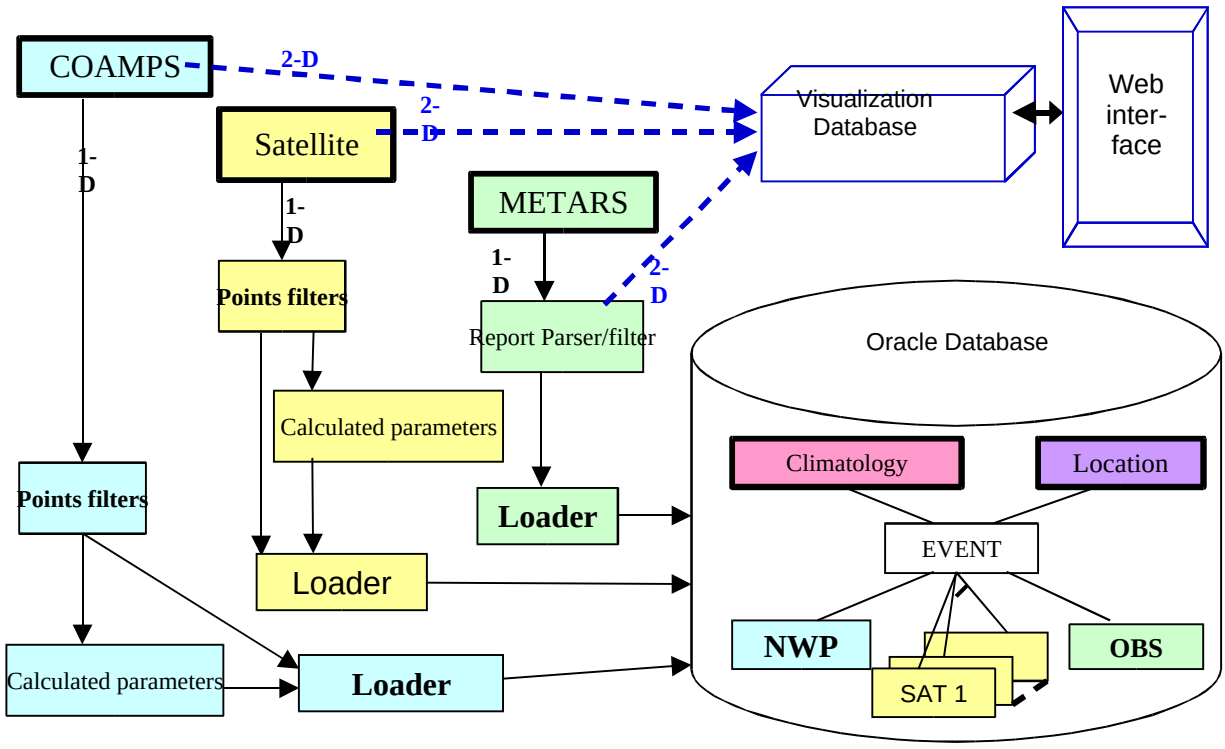

Figure 4 Data flow and schema. The three continuous data sources (COAMPS, Satellite, METARs), are stored for visualization through a web interface as two dimensional images, and are post-processed for storage in the database management system.

All parameters discussed in the previous section are computed, processed, collected, and stored in a single database. The database tables are updated once a day after all model runs and post-processing has been completed. Each table row represents the available information for a particular location at a specific hour.

The flow of data involves five steps:

1. Data generation/collection.
2. Data cleaning and pre-load processing.
3. Data loading in each individual, source-specific table.
4. Data post-load processing (calculate/update various derived fields).
5. Data consolidation: generating an *Event* record (see Figure 2) for each date/time where complete information is available (i.e., data from all three sources).

The database is organized in a star schema as shown in Figure 4. The key of each table is the day-time group and the location ID. The Climatology and Location tables are constant-valued reference tables, while the NWP, OBS, and Satellite tables are updated daily with new data, consisting of records for each specific location and day-time group..

Some pre-load steps include:

- Time rounding: adjusting the time stamp of METAR reports and satellite points to the closest hour, to correspond with the model data.
- Satellite filtering: recognizing missing data.
- Satellite derived products: low clouds, cloud optical depth, cloud classifications, environmental data records, etc.
- METAR report processing: computation of vapor pressure, cloud/no cloud, variable wind directions, etc.
- METAR cleaning: removing duplicate, later corrected, or mislabeled reports.
- COAMPS/Satellite combination products: using satellite infrared temperatures together with COAMPS profiles to determine cloud top height.

Our data mining tools require as input a denormalized (flat) table. In other words, rows representing the location of interest will be selected from all database tables containing the required information and joined together to form a single *Event* record of up to 90 variables. Each row will represent all available information for one day/time at one location.

4. METHODOLOGY

The primary tools selected for data mining are C5.0 and Cubist (Quinlan, 1992). These were selected for their ease of use and well-recognized robustness. C5.0 generates *decision trees,* which are used for classification into categories. The Cubist program creates a set of *rule-based predictive models*, which are used for regression-type estimation of continuous values. Data was extracted from the Oracle database as a flat, ASCII format file, using Oracle Discoverer. Studies were done on each focus area independently, but combining each data from all locations within each focus area. Studies examining each location independently rarely showed any improvement. This is mostly likely because of the much smaller training set available for an individual location..

The data was randomly evenly divided into training and testing sets, although 10-fold cross validation was also used for error estimates. To achieve the best results, each experiment was decomposed into three components:

1. *Determination of cloud presence*. We used C5.0 to create a decision tree which could classify each record as “Cloud” or “No Cloud” to indicate cloud presence.
2. In locations with cloud presence, d*etermination of low cloud ceiling (< 1000m) versus high cloud ceiling.* Once again we used C5.0 to create a decision tree to classify the “Cloud presence” records.
3. In locations with low cloud, *determination of cloud ceiling height* is performed using a rule set generated by Cubist.

Learning experiments were based on three different sets of variables:

1. COAMPS variables only.
2. Satellite variables only (geostationary satellite for each region).
3. Fused (combined) COAMPS and Satellite variables.

5. RESULTS

Our initial work has focused on cloud ceiling, and on the West Coast during daylight hours. Overall, the data mining method outperformed all other methods. We compared it to the COAMPS derived ceiling calculation, as well as the single source variable sets described above.

Figure 5-Figure 10 show plots of a sample location, KMRY, in several variations. Figure 5

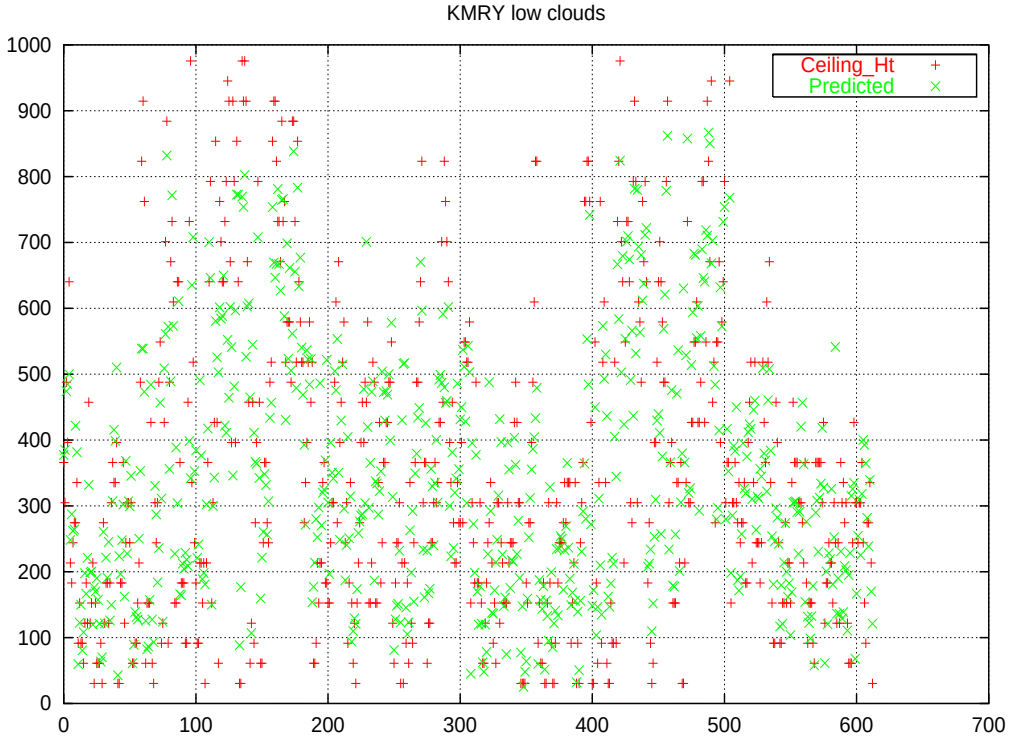

Figure 5. Plot of data mining-predicted low cloud ceiling heights ("Predicted") vs. ground truth ("Ceiling_Ht").

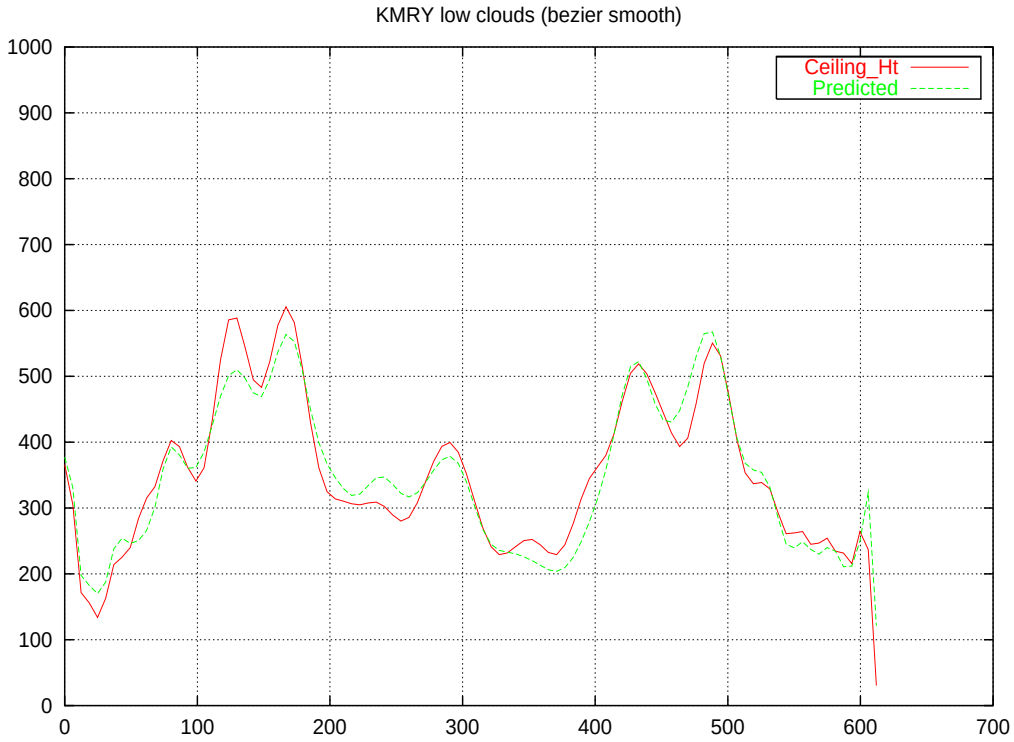

Figure 6. Same as previous figure, but with Bezier smoothing of datapoints.

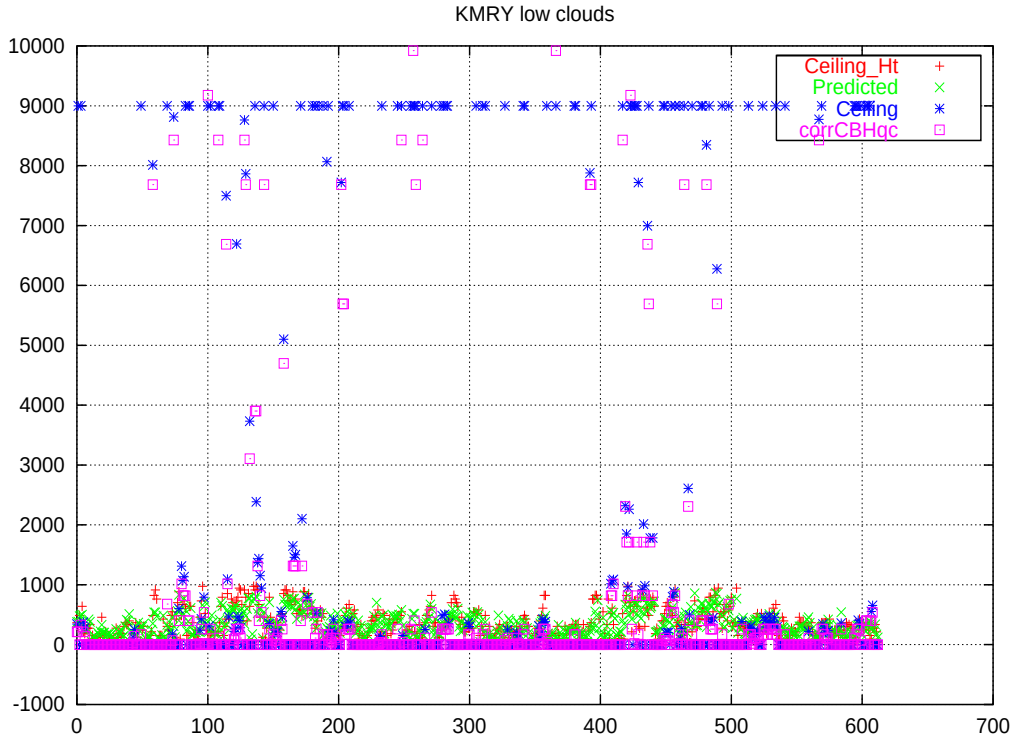

Figure 7. Plots of observed ceiling ("Ceiling_Ht"), data mining predicted ceiling ("Predicted"), COAMPS derived ceiling parameter ("Ceiling"), and COAMPS water mixing ratio derived cloud base height ("corrCBHqc").

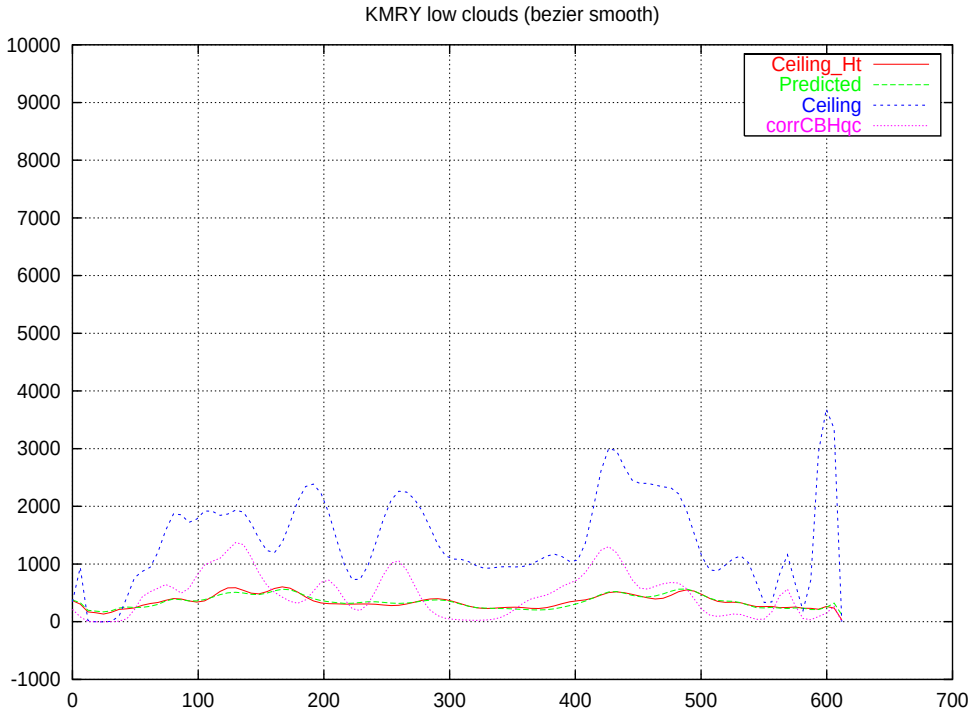

Figure 8. Same figure as previous, but with Bezier smoothing.

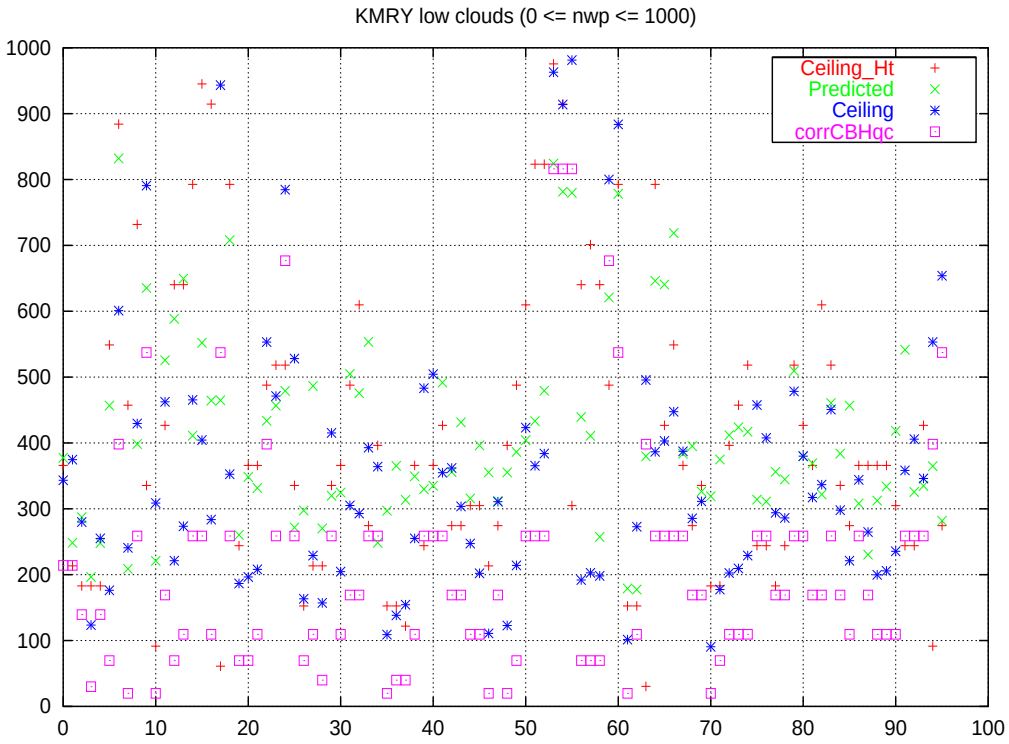

Figure 9. All plots, with only cases where COAMPS ceiling < 1000m..

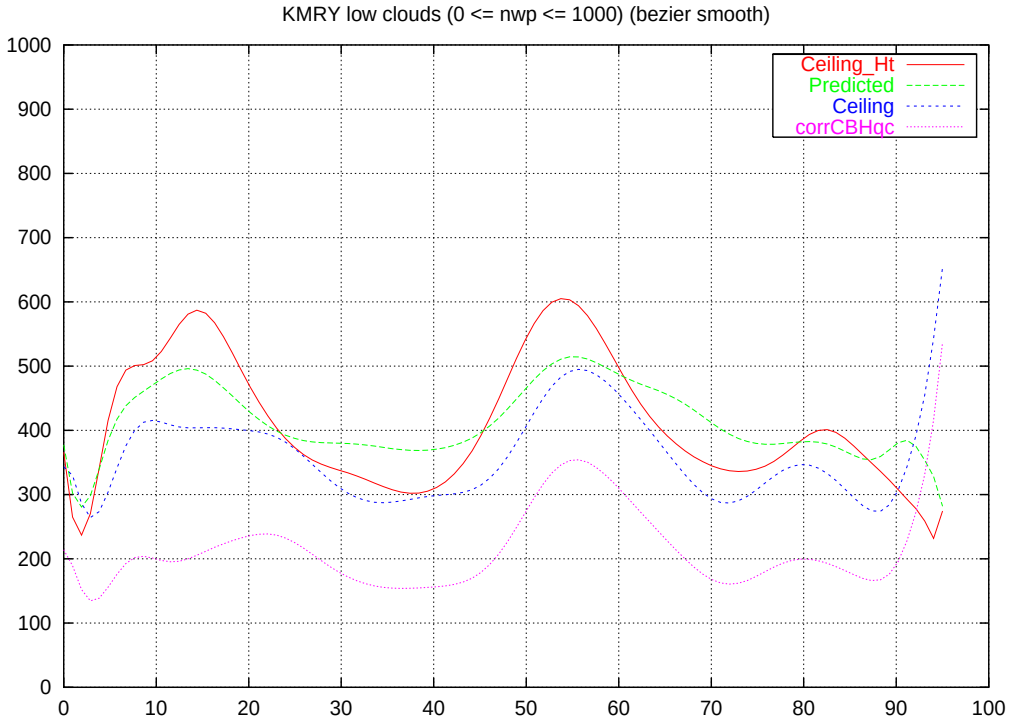

Figure 10. Same as previous plot, with Bezier smoothing..

shows a comparison of our predicted ceiling height for low clouds ("Predicted"), versus the true observed ceiling height ("Ceiling_Ht"). Figure 6 reflects the same data, but smoothed using a Bezier technique. The smoothed plots show a clear relationship between the cloud ceiling we have calculated, and the ground truth values. The next two figures show the same plot, but with COAMPS ceiling estimates included, both the COAMPS derived ceiling, and a cloud base derived using the water mixing ratio. Note that, because COAMPS has difficulty determining low cloud ceiling cases, there are a great many cases plotted which are actually high cloud ceiling, or no ceiling captured in that data set. The smoothed plot shows this bias clearly. The final pair of figures plot only that data where both COAMPS and our method indicate low ceiling. Once again, the smoothed plot shows a close correspondence between our method and ground truth, and indicates a bias in the COAMPS estimates. Note that all the data was randomly shuffled and replotted, resulting in a similar correspondence in Bezier-smoothed plots, suggesting that the degree of correspondence is independent of any temporal information in the data.

Table 3 Shows error comparisons for the entire west coast. The COAMPS ceiling method is the derived calculation produced by COAMPS. The remaining three methods are all created using C5.0/Cubist, on different sets of variables: COAMPS variables only (except the ceiling product), satellite variables only, and the fused set of variables. It is clear from the results, the data mining method outperforms COAMPS in average error, and the benefit of using fused COAMPS and satellite variables is demonstrated.

West Coast Locations	Error (%)		Avg. error
			correlation
Method	Cloud presence	Low cloud detection	Cloud base height
COAMPS ceiling	25.5	50.2	213.0
			0.5
COAMPS variables	21.0	21.9	128
			.73
Satellite variables	12.2	22.2	170
			.52
Fused variable set	**11.1**	**19.9**	**126**
			.74

Table 3 Classification and regression error comparisons

6. CONCLUSION AND FUTURE WORK

Our initial results demonstrate the viability of using KDD to discover algorithms which can locate cloud presence, and calculate cloud ceiling more successfully than numerical weather prediction models. Furthermore, results indicate that both COAMPS and satellite variables make contributions to the final results - - indicating the value of a "fused data" approach. Still remaining is a study of night-time hours, and the other focus areas. Also, we are still in the process of applying these methods to rain rate and rain accumulation.

ACKNOWLEDGEMENTS

The support of the sponsor, the Office of Naval Research, under Program Element 0602435N is gratefully acknowledged. The assistance of Jeff Hawkins, John Cook, Mike Neith, the satellite group, and COAMPS/TAMS-RT developers (all with NRL Monterey), and Melanie Wetzel with the Desert Research Institute is greatly appreciated.

REFERENCES

Fayyad, U., G. Piatetsky-Shapiro, P. Smith, R. Uthurusamy, eds., 1996: *Advances in Knowledge Discovery and Data Mining.* AAAI Press/MIT Press, Menlo Park, CA.

Hadjimichael, M., P.M. Tag, R.L. Bankert, 1998: Discovering Model Bias in the Determination of Cloud Base Height. *Proceedings, First American Meteorological Society Conference on Artificial Intelligence*, Phoenix, AZ, January 11-16, 1998, J5--J8.

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

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

Quinlan, J., 1992: *C4.5: Programs for Machine Learning,* Morgan Kaufmann Pub., San Mateo, CA.

Turk, J., Liou, C.S., Qiu, S., Scofield, R., Ba, M.B., and Gruber, A., 2001: Capabilities and characteristics of rainfall estimates from geostationary- and geostationary+microwave-based satellite techniques. *Proceedings of Symposium on Precipitation Extremes: Prediction, Impacts, and Responses*, Amer. Meteor. Soc., 191-194.

Weiss, S.M. and Indurkhya, N., 1998*: Predictive Data Mining, a practical guide.* Morgan Kaufmann Pubs, Inc., San Francisco.

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