NEW TECHNIQUE FOR CLOUD MODEL CONTROLLED PRECIPITATION RETRIEVAL:

CLOUD DYNAMICS AND RADIATION DATABASE (CDRD)

P 1.21 DATA MINING APPLICATIONS AT GLOBAL AND REGIONAL SCALES

[1]Joseph Hoch *, [2]C.M. Medaglia, [3]A.V. Mehta, [2]A. Mugnai, [3]E.A. Smith, [1]G.J. Tripoli

[1] University of Wisconsin, Madison, Wisconsin
[2] Institute of Atmospheric Sciences and Climate – Rome, Italy
[3] Goddard Space Flight Center – Greenbelt, Maryland

1. INTRODUCTION

Passive microwave remote sensing of precipitation from platforms such as the Special Sensor Microwave Imager (SSM/I), the Advanced Microwave Scanning Radiometer (AMSR), and the Tropical Rainfall Measurement Mission (TRMM) has been a major focus in hydrological research for the past several years. Successful estimation of precipitation from these platforms relies on the accuracy of the particular retrieval algorithm being utilized. Retrieval algorithms are based on cloud radiation databases (CRDs) to relate in-situ measurements of brightness temperatures and radar reflectivity profiles to a-priori microphysical profiles found in the CRDs. One problem with CRD retrieval based systems is that profiles can be chosen that are unrepresentative of the dynamical and thermodynamical state of the atmosphere. We have recently introduced the concept of the Cloud Dynamics and Radiation Database (CDRD) for precipitation retrieval purposes. The CDRD concept is an improved version of the current CRDs. The CDRD contains the same information as the present CRDs, but in addition contains information about the dynamical and thermodynamical structure of the atmosphere.

The CDRD contains dynamical tags that are computed from a cloud-resolving model, the same simulations used to calculate brightness and microphysical profile information. The CDRD also provides the opportunity to investigate relationships between the microphysical structure of precipitation systems to large-scale dynamic and thermodynamic variables of the atmosphere.

* *Corresponding author address:* Joseph Hoch, Univ. of Wisconsin, Dept of Atmospheric and Oceanic Sciences, Madison, WI 53715; email: hoch@wisc.edu

The objective of this paper is to first discuss the methodology for implementing the CDRD version 1.0 system over the globe and showing the possibility to retrieve useful subsets of information from a massive global database system. The dynamical and thermodynamic tags available in the CDRD are presented. The primary technique for extracting useful subsets of information is through the use of data mining techniques. Section 2 discusses several data mining techniques. Section 3 of the paper focuses on the implementation of the CDRD design. Section 4 focuses on the application of the CDRD system for data mining purposes. A sample CDRD database has been designed with eight different cloud resolving model (CRM) simulations. This test database is used to highlight data mining techniques for global and regional applications. A regional case, over central and southern California, focuses on orographically enhanced precipitation. This section specifically focuses on using the CDRD for hypothesis testing. The following hypothesis is purposed and tested: as dynamical tag dimensions increase the variance properties for the acquired profiles decreases.

2. DATA MINING

Simply defined, data mining is the use of data analysis tools to discover unknown relationships from large data sets. Data mining techniques are often used to predict future trends and behaviors allowing for knowledge-based decisions. Data mining, growing in popularity, is used in both the public and private sectors. Industries such as banking, insurance, medicine, and general business often use such tools to reduce costs and raise efficiency. As the amount of data available to the earth science community continues to increase rapidly, data

mining has become a more desirable research tool to utilize large databases effectively.

Data mining techniques are utilized as a tool for efficiently extracting useful subsets of information from the CDRD. With the use of such techniques previously unknown relationships between dynamical/ thermodynamic tags and microphysical properties can possibly be observed. Managing the physical size of the CDRD data warehouse presents a challenge of extracting useful information. Advanced data mining techniques are used to effectively retrieve appropriate parameters for retrieval purposes. Possible data mining techniques are now discussed in further detail.

Table 1 provides a short summary of popular data mining processes. Neural networks are analytic techniques that are modeled after the cognitive process of learning and the neurological functions of the brain. This method is capable of predicting new observations from previous observations. The technique is able to *"learn"* from existing data already present in the data warehouse. Denby discusses how artificial neural networks are used in high-energy physics research (1993).

A decision tree is a model that is both predictive and descriptive (Frank and Whitten, 2005). It is called a decision tree because the resulting model is presented in the form of a tree structure. The visual structure of a decision tree makes the tool easy to understand. Decision trees are most commonly used for classification by predicting what group a case belongs to. Decision trees can also be used for regression, predicting a specific value. The primary output from a decision tree algorithm is the tree itself. The training algorithm that creates the tree is referred to as induction. Decision trees are commonly used in business to make decisions, based on *if-then* relationships.

Genetic algorithms are similar to the process of natural selection. This method searches for the most optimal matches based on certain criteria or combinations. The desired quantity ("organism") is retrieved based on the most optimal set of criteria ("genes"). Genetic algorithms have been used in finance but are not very practical as a data analysis tool. This is due to the lack of statistical significance from the obtained solution.

The nearest neighbor method, or sometimes referred to as the k-nearest neighbor method, refers to similar data points, within a data warehouse, that are "living" in each other's neighborhood. The "k" refers to the number of "neighbors" being investigated to retrieve a certain quantity. For example, 6-nearest neighbor looks at six neighbors. This data mining tool is more of a search technique than a learning tool. This technique is often used with small subsets of data.

Finally, rule induction is a data mining technique that extracts statistically significant data using if-then rules. Cohen (1995) presents a fast effective example of a rule induction technique. This tool can be used to infer generalizations from the information in the data.

A rule induction scheme is the best option to effectively mine data from the CDRD, based on the statistical properties of the scheme. The proposed data-mining scheme allows for the selection of microphysical profiles from the database using dynamical and thermodynamical variables linked to microphysical profiles. The available variables ("tags") in the CDRD are discussed later in the paper. The CDRD mining algorithm retrieves the appropriate profiles and computes the corresponding variance, throughout the entire vertical atmospheric column. The matching of microphysical profiles with dynamic tags relies on a Bayesian selection approach.

The advantage of using dynamic and thermodynamic variables for microwave remote sensing is because such tags increase the number of constraints on the retrieval algorithms. In a Bayesian selection scheme, when limited by more constraints, the database should provide increased representative profiles for a particular precipitation system.

3. CDRD IMPLEMENTATION

.

The fundamental core of the CDRD system is based around a CRM. The UW-NMS is used to produce simulations for the formulation of the CDRD version 1.0 system. The UW-NMS is described in detail by Tripoli (1992). The model used for radiative transfer calculations is the Successive Order of Interaction (SOI) Radiative Transfer Model. The SOI is a one-dimensional azimuthally averaged, plane-parallel radiative transfer model. This model includes the effects of scattering from all hydrometeors. Atmospheric polarization is ignored, but not surface polarization (Heiginger, O'Dell, Bennartz, and Greenwald 2005).

Every day a random global location is selected for a new CRM simulation. The random locations move between four global regions, based on the equator and international dateline. This technique is used for somewhat equal global simulation spacing. Figure 1 shows where the random inner grids are located for the

sample database (two simulations per region). This type of image is updated daily and available online. The UW-NMS is used to simulate a 12 hour prediction over the selected location. Microphysical profiles and dynamical/ thermodynamical tags are saved at the 12 hour forecast time. Vertical profiles, at all 36 levels, of microphysical variables are saved based on simulated surface precipitation rates. The criteria for saving a profile in the database occurs when surface rain rates are 0.50 mm hr^{-1} or greater and/or frozen (snow, graupel, aggregates, pristine crystals) surface rates are 0.25 mm hr^{-1} or greater. These criteria were selected based on the capability of current microwave remote precipitation sensors. These precipitation criteria are near the accepted lower limits of useful satellite data. A sample grid setup, over the California region, is shown in figure 2. Table 2 outlines the specific variables that are included in a "microphysical profile". The available dynamic and thermodynamic tags which are paired with microphysical profile points in the CDRD system are listed in tables 3 and 4. Table 3 lists the variables that are produced from the outer grid of the UW-NMS, at 50km resolution. These variables are referred to as large-scale tags. Table 4 lists the variables that are produced from the inner grid of the UW-NMS, at 2km resolution. These variables are referred to as the high-resolution tags.

4. CDRD APPLICATION

The following section shows how the CDRD can be utilized to retrieve the "best-possible" microphysical profile for a particular event. Of particular focus in this paper is the severe storm that impacted California, from January 7th - 11th, 2005. This storm brought heavy orographic precipitation over much of the Sierra Nevada mountain chain. Certain areas in California recorded over 25 inches of equivalent rainfall. The case-study retrieval time is January 8th, 2005 at 12Z over the selected domain. Figure 3 shows the accuracy of the UW-NMS simulation. Accumulated precipitation from January 7th – 11th is compared with NCEP stage IV radar data.

At first, microphysical profiles are selected from the 8-run sample CDRD database using only brightness temperatures, at 89.0 GHz. Figure 4 shows a simulated 89.0 GHz field, produced by the UW-NMS and SOI models. The idealized simulated TB field is taken as "truth" for selection of microphysical profiles from the CDRD. The goal is to retrieve the best profile for the Sierra Nevada region, where the majority of orographic precipitation is occurring. The simulated overpass suggests microphysical profiles should be selected from a range of brightness temperatures (180 - 240K). These are the brightness temperatures that are occurring over the mountain range.

Using the CDRD with only brightness temperatures is similar to a CRD approach. Next, profiles are selected using brightness temperatures paired with certain dynamical and thermodynamical tags. First, microphysical profiles are selected using the brightness temperatures along with mean sea level pressure and surface temperature. Second, profiles are selected using 89.0 GHz brightness temperatures, mean sea level pressure, surface temperature, and topography elevation. Since the database being used only contains 8 separate model runs, over the entire globe, three extra tags along with brightness temperatures is enough to correctly identify the desired California profiles.

Table 5 shows the range for each tag used and the number of microphysical profiles that are retrieved. Notice that the number of possible microphysical tags decreases around 91 percent by adding only two dynamical tags. The tags were taken from the outermost grid of the UW-NMS, which uses 50km grid spacing. The follow-up paper uses an operational model such as the GFS forecasting system to obtain the required dynamical tags.

The speculated hypothesis for showing the advantages of the CDRD approach is that the variance of microphysical profiles decreases as the number of dynamical tags increases. Figure 5 shows the mean vertical profile for total condensate mixing ratio. The mean changes as the number of profiles decreases. Figure 6 shows the structure of the variance in the microphysical profiles. When using only 89.0 GHz brightness temperatures, there is a significant amount of variance in the retrieved microphysical profiles. When dynamical tags are included, the variance of the retrieved profiles drops significantly. When using the CDRD.v1 system for real-time applications it is theorized that more than two or three dynamical/ thermodynamical tags are needed to narrow the search for the "true" microphysical profile. In this case, since there are only 8 runs over the globe, the addition of the first two tags narrowed the focus to the California region. Eventually, the CDRD will be made up of hundreds of simulations globally.

5. CONCULSIONS

This paper highlights several data mining techniques and applications. A rule induction scheme best matches the technique used for retrieving information from the CDRD data warehouse. A sample CDRD database, 8 simulations, is used to highlight the benefit of using the CDRD approach rather than a traditional CRD approach. An orographic case study over California is used to show the value added, by using dynamic and thermodynamical tags along with brightness temperature fields. Because of these tags, the average vertical variance structure in the profiles retrieved significantly decreases. This paper has shown that by using the CDRD tag approach more accurate microphysical profiles can be retrieved from the accompanying database.

The CDRD is a robust system that improves microwave precipitation retrieval techniques. This system can also be used for many other earth science applications. The CDRD is available online and can be used to investigate many possible relationships between microphysical quantities and atmospheric parameters.

6. REFERENCES

Cohen W. W., 1995: Fast effective rule induction. In Machine Learning, *12th International Conference,* Morgan Kaufmann.

Denby B., 1993: The Use of Neural Networks in High Energy Physics, *Neural Computation*, **5**, num.4, 505-549.

Heiginger A. K., C. O'Dell, R. Bennartz, T. Greenwald, 2005: The Successive Order of Interaction Radiative Transfer Model, Part I: Model Development, *J. Atmos. Sci*., submitted.

Tripoli G. J., 1992: A Nonhydrostatic Mesoscale Model Designed to Simulate Scale Interaction, *Mon. Wea. Rev.*, **120**, 1342-1359.

Whitten I. H., E. Frank, 2005: Data Mining: Practical Machine Learning Tools & Techniques with Java implementations, **2nd ed**., Morgan Kaufmann.

ACKNOWLEDGEMENTS

The authors would like to thank Ralf Bennartz, Mark Kulie and Chris O'Dell with their help on the SOI radiative transfer model.

TABLE 1 – Popular Data Mining Processes

Artificial neural networks	Non-linear predictive models that learn through training and resemble biological neural networks in structure.
Decision trees	Tree shaped structures that represent sets of decisions. These decisions generate rules for the classification of the dataset.
Genetic algorithms	Optimization techniques that use processes such as combination.
Nearest neighbor method	A technique that classifies each record in a dataset based on a combination of classes of the k record(s) most similar to it in a historical dataset.
Rule Induction	The extraction of useful if-then rules from data based on statistical significance.

Table 1: Possible data mining techniques for CDRD data mining

TABLE 2 – UW-NMS Microphysical Profile

Total Condensate Mixing Ratio
Rain Mixing Ratio
Cloud Mixing Ratio
Water Vapor Mixing Ratio
Graupel Mixing Ratio
Aggregate Mixing Ratio
Pristine Crystal Mixing Ratio
Surface Precipitation Rates (Rain, Snow, Aggregate, Pristine Crystal, Graupel)
Surface Skin Temperature
Q1, Q2
Temperature
Pressure
Height
Zonal Wind (U)
Meridional Wind (V)
Vertical Velocity (w)

Table 2: CDRD Microphysical profile variables

TABLE 3 – Large Scale Dynamic Tags

Mean Sea Level Pressure (hPa)	Freezing Level (m)	Surface Theta Gradient	LFC Height (m)
Surface Temperature (F)	Lifted Index	700mb Theta Gradient	LCL Height (m)
**U-Wind (m/s)	Froude Number	Surface Theta-E Gradient	Topography Height (m)
**V-Wind (m/s)	Surface Theta-E (K)	700mb Theta-E Gradient	PBL Height (m)
U Momentum Flux	Surface Brunt Vaisala Frequency	** Q Vector Convergence	Richardson Number in the PBL
V Momentum Flux	**Temperature	Surface Divergence	Potential Vorticity Advection at 700 and 250 mb
CIN (J/kg)	Potential Vorticity at 700 and 200 mb	Divergence at 700 and 200 mb	Height of Maximum Cape (m)
Maximum Cape (J/kg)	Surface Vertical Vorticity	**Vertical Velocity (m/s)	Diabatic Moisture Term
Surface Cape (J/kg)	Vertical Vorticity at 700 and 200 mb	Theta-E minimum (K)	Latent Heat Term
Kinetic Energy	0-6km Wind Shear	500 and 850 mb thickness (m)	**Specific Humidity

Table 3: Large-scale tags for the CDRD.

** Denotes vector variables (1000,925,850,700,500,250,200,150,100mb)

TABLE 4 – High-Resolution Dynamic Tags

Cloud Ceiling (m)	** Temperature (K)
Topography Height (m)	** Q Vector Convergence
Largest Topography Neighbor Difference (m)	** Vertical Velocity (m/s)
Direction of Topography Direction (degrees)	Cloud Fraction
PBL Height	Convective Cloud Fraction
Mean Sea Level Pressure (hPA)	Stratiform Cloud Fraction
Surface Pressure (hPA)	

Table 4: High-resolution tags for the CDRD.
** Denotes vector variables (1000,925,850,700,500,250,200,150,100mb)

TABLE 5 – CDRD Data Mining Case Study –
California Orographic Precipitation

Parameter	Range	Number of
89.0 GHz Temps (K)	*180 - 240*	*132474*
Mean Sea Level Pressure (hPA)	*1008 - 1014*	*6138*
Surface Temperature (K)	*262 - 282*	*4990*
Elevation (m)	*700 - 4000*	

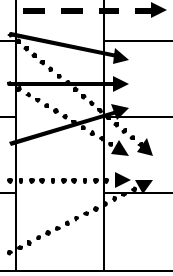
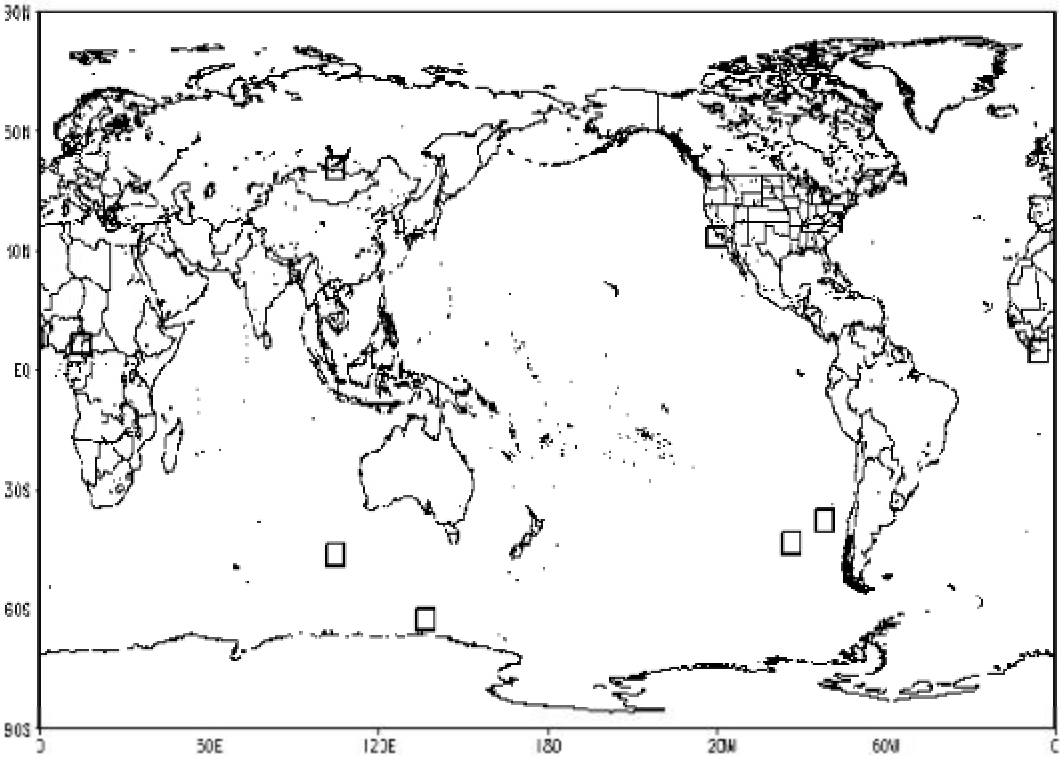


Table 5: Statistics from the data mining orographic case study



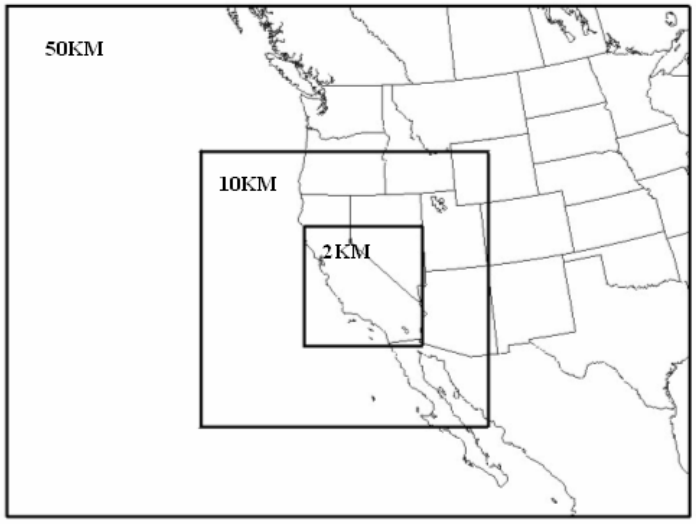


Figure 1. Sample CDRD database CRM locations

Figure 2. Nested grid structure used for UW-NMS CDRD simulations. This figure shows the setup over the California case study region

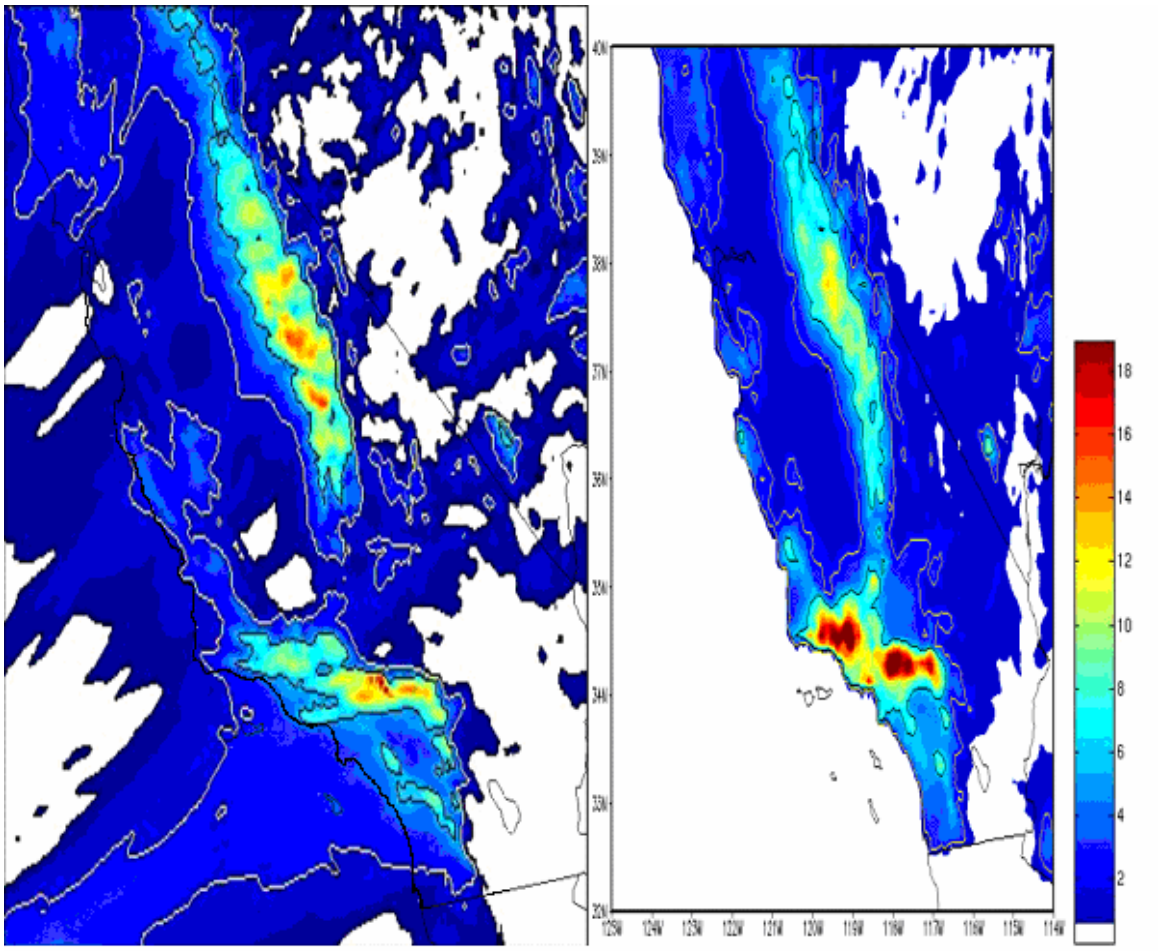


Figure 3. Left: 4 Day Accumulated Rainfall (in) (12Z Jan 07 – 12Z Jan 10) from UW-NMS
Right: 4 Day Accumulated Rainfall (in) (12Z Jan 07 – 12Z Jan 10) from NCEP Stage 4 Radar Data
Yellow Contour – 3 inches
Black Contour – 7 inches

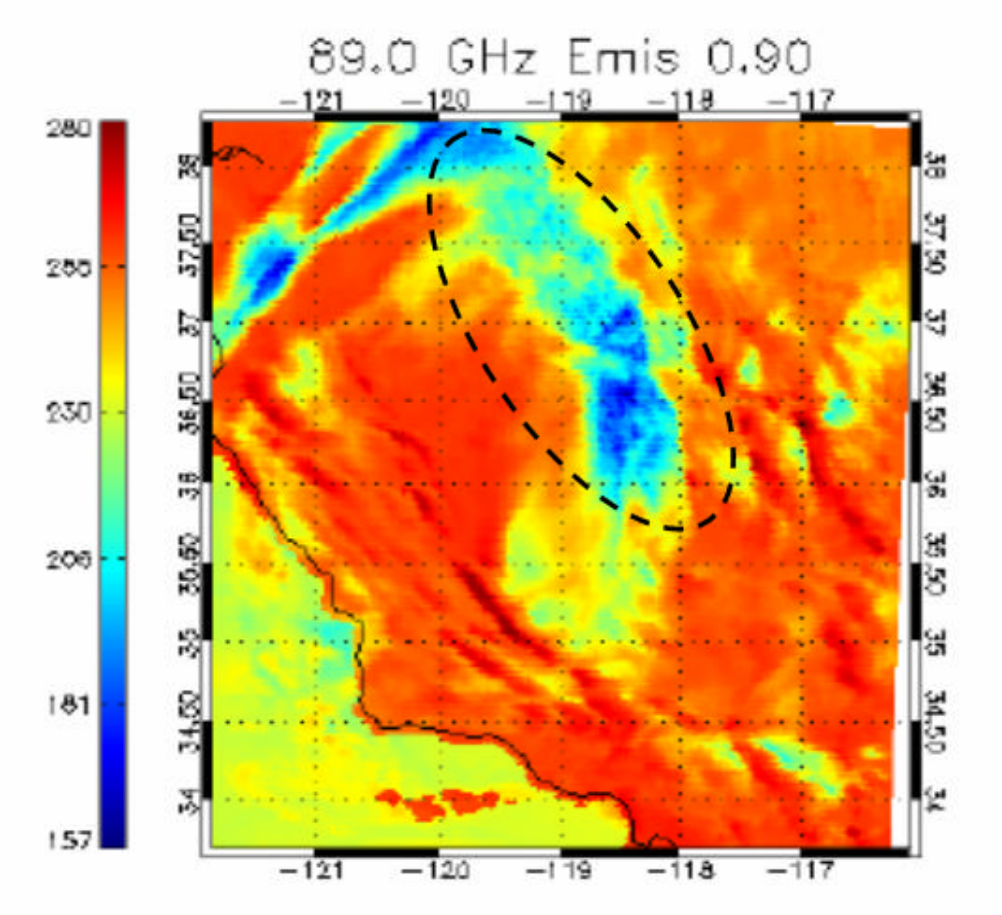


Figure 4. This figure shows the simulated 89.0 GHz brightness temperature field for the California case study region on January 8th, 2005 at 12Z.

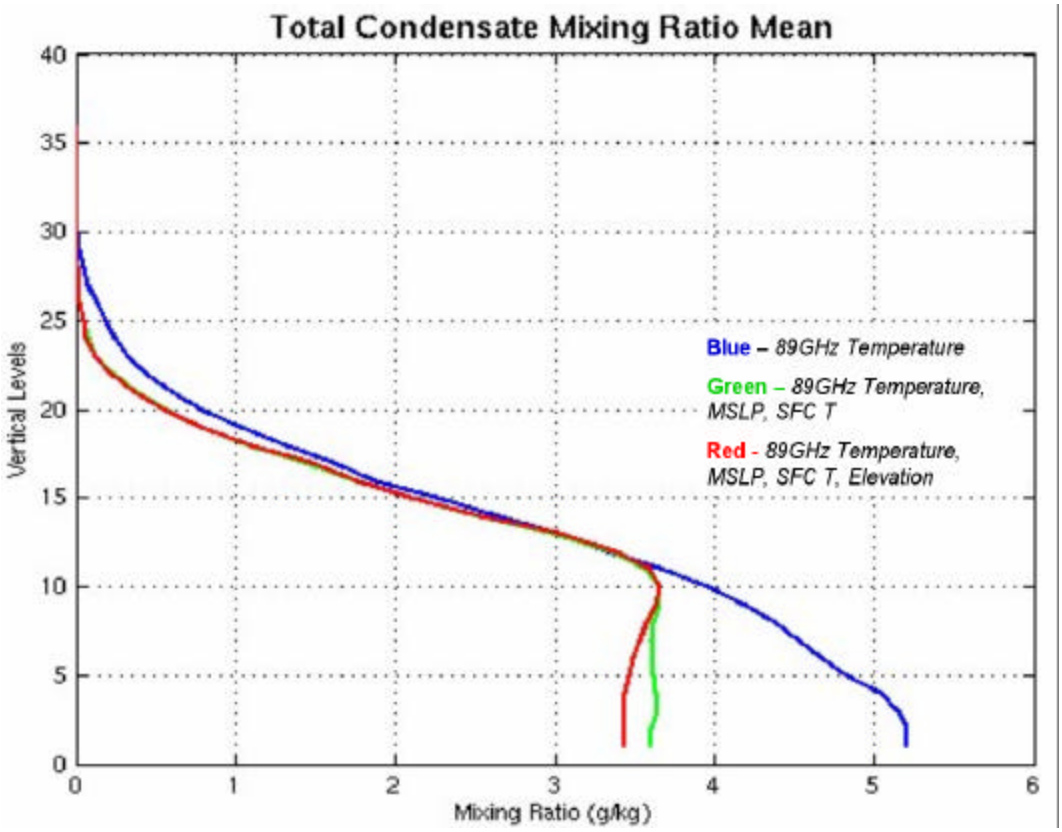


Figure 5. This figure shows the total condensate mixing ratio mean for the three different types of retrieval.

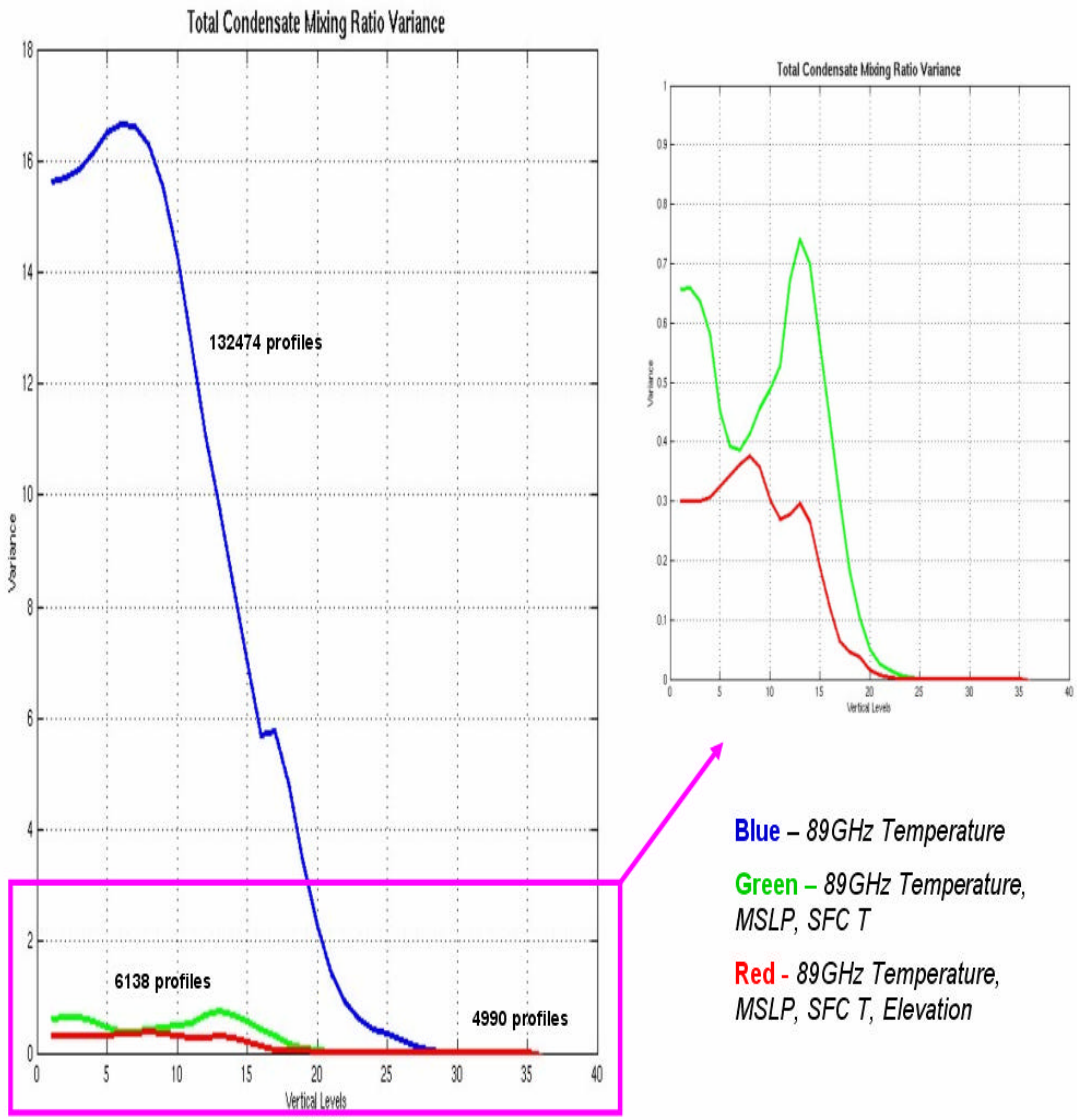


Figure 6. This figure shows the total condensate mixing ratio variance taken in the vertical for the three different types of retrieval.