N96-12915

p, 8

MOM: A Meteorological Data Checking Expert System in CLIPS

Richard O'Donnell
Geophysics Laboratory
Hanscom AFB, MA

ABSTRACT

Meteorologists have long faced the problem of verifying the data they use. Experience shows that there is a sizable number of errors in the data reported by meteorological observers. This is unacceptable for computer forecast models, which depend on accurate data for accurate results. Most errors that occur in meteorological data are obvious to the meteorologist, but time constraints prevent hand-checking. For this reason, it is necessary to have a "front end" to the computer model to ensure the accuracy of input. Various approaches to automatic data quality control have been developed by several groups.

MOM is a rule-based system implemented in CLIPS and utilizing "consistency checks" and "range checks". The system is generic in the sense that it "knows" some meteorological principles, regardless of specific station characteristics. Specific constraints kept as CLIPS facts in a seperate file provide for system flexibility. Preliminary results show that the expert system has detected some inconsistencies not noticed by a local expert.

I. Introduction

Large amounts of meteorological data must be processed in order to study and forecast our weather. The accuracy and utility of forecasting models and techniques depend heavily on the accuracy of the input data.

At the Geophysics Laboratory, Hanscom Air Force Base, in Bedford, Massachusetts, there is a meteorological data collection facility called AIMS (Air Force Interactive Meteorological System). AIMS is a VAXcluster with many sources of automated continual data input, including the FAA 604 line, and a GOES ground station, which supplies satellite imagery and data. (FAA stands for the Federal Aviation

Administration, which oversees many flight related forecasting operations. GOES is an acronym for Geostationary Operational Environmental Satellite. The GOES ground station supplies satellite imagery for the Western Hemisphere every 30 minutes.) The FÅA 604 line includes Service-A data (hourly data within North America), synoptic data (data from worldwide sources every six hours), radar data, forecasting model results, and other types of data.

The purpose of this facility is to develop new techniques to study and forecast atmospheric behavior. The forecasting models being developed use these sources of data as input to generate a forecast. It is obvious that this data needs to be accurate in order for these models to provide accurate output. This is the problem of meteorological data validation. One purpose of our study was to determine exactly how frequently inaccurate observations are reported.

II. Meteorological Data Validation

The errors in meteorological data are produced by two sources: human error, and machine error. Human errors could entail a misreading of an instrument, a mismeasurement, or even a "typo", while machine errors include malfunction, breaking of equipment, and even noise in the data lines. All of these factors combine to cause data available to the scientists and the computer models to be in error.

The Meteorological Observation Monitor (MOM) is an attempt to weed out errors in the database by identifying errors that are found. MOM is written in CLIPS and is still in the process of being tested and further developed.

MOM is a system made up of four basic parts: a main knowledge base of CLIPS rules, a base of specific meteorological facts, a module which extracts the data from the database and puts the data into the form of CLIPS facts, and, of course, the database itself. The main knowledge base and the meteorological fact base are the components to be studied, since they represent the expert system part of MOM, and are the parts written in CLIPS.

In order that MOM be made more flexible and expandable, as well as maintainable, only general priciples were included in the main knowledge base, and specific meteorological information was left out. The specific data needed to make decisions was included in the fact base. For example, the main rule base contains the general information that there is a minimum air temperature at which rain may occur. The specific temperature that will be used to determine whether the type of precipitation is correct resides in the fact base. This modular design lends itself well to maintenance, especially since data is sometimes invalid because it does not conform to reporting conventions, and these conventions can change. For example, wind gusts speed may not be reported unless the gusts of wind are at least 10 knots greater than the low wind speed for the hour. This convention has changed through the years, and it is possible it will change again. Updating this type of information would require only minimal maintenance to MOM, since only the smaller fact base would need to be changed.

Before getting any further into the design of MOM, it would be best to discuss what specific problems arise with meteorological data, and several methods to validate data. It may seem obvious, but meteorological data is invalid whenever it does not accurately represent the real world. Choosing which of these data are accurate, and which are not, is not always possible. In many cases, however, situations arise which clearly show the existence of invalid data. For example, the temperature at Logan Airport in Boston may truly be 64°F, but is being reported as 69°F. To the scientist sitting in the lab in Bedford, 69°F seems well within the realm of possibility, and that data will never be found to be invalid. This is not catastrophic, because if this kind of invalid data goes unnoticed, it is not very disruptive to the computer models that produce forecasts. However, there are times when the scientist in the lab may know for certain that the data is invalid, if Logan is reporting 75°F in January, for example.

There are in principal two reasons data can be invalid. First of all, it can break physical laws of nature. Rain is highly improbable when it is 5°F. The other reason reported data can be invalid is that it can break conventions, such as the wind gusts convention mentioned earlier. While it may be true that winds are from the north at 6 knots with gusts up to 9 knots, to report that is not helpful, and would cause others to question the validity of the data, since it is not possible for the difference from lull to peak wind to be 10 knots. There are several such conventions, and we will see some of these later.

There are at least four different methods one may use to successfully recognize invalid data. The first method, and the one most often used by a human meteoroligist scanning the weather maps, is "buddy" checks: that is, checking the nearest neighbors of the station that is reporting the data to validate it. If Boston is reporting 14°F, and Bedford 50°F, there is an enormous discrepancy to account for. The second method is to do a time check. If New York's Kennedy Airport is reporting a temperature that is in question, a time check would look at the most recent reports of temperature at Kennedy and compare them to the data in question. The third method, and the primary method employed by MOM, is to do a consistency check. A consistency check takes an hourly report consisting of several observed parameters, and determines whether the relationships between the parameters are consistent. For example, if a station reports a temperature of 50°F but also reports snow, there is an inconsistency in the report. A fourth method of validation is to do a range check. A range check takes a single data item and determines whether it falls within climatological extremes for the reporting station and month. A primitive range checker is also included in MOM. A complete system would use all of these methods to best validate data.

There are problems with each of these methods. Some of these problems are meteorological and some are computational. The buddy checking method has a problem in that each station would need to have buddies, and not all have near neighbors. Not only is that the case, but sometimes, because of geographical elements, a nearby neighbor would not be as good a choice as a further neighbor. Therefore, a table of neighbors would need to be created so that only those neighbors which would contribute similar data would be consulted. Time checking also has problems, primarily meteorological. In many places, drastic changes in temperature can take place within an hours time, which is the normal reporting interval. These drastic changes may be extremely improbable elsewhere. Self-consistency checks have the problem of being too limited. The data may not disagree; however, that does not necessarily indicate that there are no errors present. Range checking is similar; if data is flagged for being out of a reasonable range, it is a good bet it is invalid, but alot of invalid data meets the requirements of that test, and therefore is not discovered. Any one method alone will not discover all errors present in the data.

III. MOM and Validation of Data: Consistency and Range Checks

When the problem of data validation was first considered in this study, it was decided that MOM would represent a first attempt to address the concern. The data chosen to be validated was Service-A data, and MOM was to employ consistency checks and range checks on this data. The reason consistency checks were selected was that, of all the methods described, it lends itself most handily to a "rules" oriented knowledge base.

Service-A data is hourly data reported from all stations in North America. MOM examines nine parameters in a report for self-consistency: air temperature, dewpoint temperature, pressure, altimeter setting, wind speed, wind gust speed, wind direction, visibility, and current weather. Pressure is not reported from a number of smaller airfields, and instead, these stations only report an altimeter setting. Except in one case, each of these parameters is a floating point number which is defined by a specific range of possibilities. For example, the range of wind direction is 0.0 to 360.0. The exception is current weather. This is defined by a string of characters, each representing a different weather pattern or phenomenon. If nothing is currently happening at a particular station, the current weather string is empty. Examples of current weather are fog and rain. A complete list of possibilities is given in Table 1.

Current Weather		
Obstructions to Vision	Weather Symbols	Intensity Symbols
F = Fog	T = Thunderstorm	+ = heavy
GF = Ground Fog	L = Drizzle	- = light
IF = Ice Fog	ZL = Freezing Drizzle	W = showers
D = Dust	R = Rain	
K = Smoke	ZR = Freezing Rain	no modifier indicates moderate intensity
H = Haze	S = Snow	
BD = Blowing Dust	SP = Snow Pellets	
BN = Blowing Sand	SG = Snow Grains	A and IC have no intensity symbols
BS = Blowing Snow	IP = Ice Pellets	
BY = Blowing Spray	IC = Ice Crystals	
	A = Hail	T may only have +

Table 1: Reporting codes for current weather

The representative letters in Table 1 can be combined in many ways, with precipitation types coming first, and obstructions to vision last, to describe the wide variety of possible weather conditions. The intensity symbols are modifiers that add to the meaning of the character preceding them. For example, the string "RF" means the reporting station is experiencing both rain and fog, while "R-F" means the station is experiencing light rain and fog. Intensity symbols are not used with obstructions to vision. These strings can be arbitrarily long to describe very mixed kinds of weather, like the weather we get in New England. On an unusually bleak winter day, a report could be "ZL-ZR-S-F" which means a mix of light freezing drizzle, light freezing rain, light snow, and fog. "TRW" means thunderstorms and rain showers. A problem with this system is that strings can be ambiguous. For example, the string "SGF" could mean either snow with ground fog, or snow grains with fog.

The nine data items discussed have many different interrelations that force a large number of rules governing consistency checking between the parameters. Table 2 shows which parameters are closely related.

Parameter	Related Parameter
temperature	dewpoint temperature
temperature	current weather
dewpoint depression*	current weather
visibility	current weather
wind speed	wind gusts speed
wind speed	current weather
pressure	current weather
pressure	altimeter setting
altimeter setting	current weather

* dewpoint depression is temperature minus dewpoint temperature

Table 2: Reported parameters which have relationships to each other

As you can see, current weather is the most commonly related parameter. Current weather is related to almost all the other parameters, and, although there are only nine distinct relationships shown in the above table, the variety in current weather forces a large number of rules. For example, the visibility relationship with current weather is just one relationship listed above. There are a large number of rules required to describe this relationship, however. For virtually every obstruction to vision and precipitation type and intensity, a rule must be created to identify the lower and upper bounds of visibility possible under the circumstances.

IV. Preliminary Results

Preliminary results show that 1 out of every 100 incoming data sets are prone to error. These results are based on close to 1200 reports that have been examined by MOM. This is a result achieved only with consistency checks. A system incorporating time and buddy checks will find many more errors. On days with mixed weather, the number of errors has been as high as 1 in 60 data items. Again, however, these results are preliminary, because most of the testing period has taken place during periods in which little or no current weather has been reported. Testing is still in process, and will continue for some time.

The majority of the errors found thus far have been reports that do not abide by conventions. A common error is reporting of wind gusts which are less than 10 knots. Another common "convention-breaking" error is a report of less than 7 miles visibility without a corresponding report of an obstruction to vision. The convention states that if a visibility less than seven miles is to be reported, an accompanying obstruction to vision must be reported.

Table 3 is an example of input to MOM. The table is a copy of a file which is read in CLIPS and processed.

```
(data station-id WORCESTER)
(data time z30-JAN-1990:11:00)
(data airtemp 86)
(data wind-dir 20)
(data wind-speed 15)
(data visibility 2)
(data current-weather freezing-rain fog)
(data precip-intensity light-freezing-rain)
```

Table 3: Sample input to MOM

The results of processing the input from Table 3 are seen in the output from MOM in Table 4.

```
CLIPS> (run)

 *** DATA FOR WORCESTER AT z30-JAN-1990:11:00 ***
     airtemp 86
     dewpt MISSING
     pressure MISSING
     altimeter MISSING
     wind-speed 15
     wind-gust MISSING
     wind-dir 20
     visibility 2
     current-weather freezing-rain fog

 *INCONSISTENT*   AIRTEMP CURRENT-WEATHER
 current weather reports freezing rain at a temper-
 ature greater than which it is likely to occur
 (max temperature for freezing drizzle is 39)

13 rules fired
Run time is 0.3203125 seconds
CLIPS>
```

Table 4: Sample output of MOM corresponding to input from Table 3

V. Future Paths of MOM

MOM is not a completed effort. Future work on MOM will be based on the outcome of testing. If work does continue on the system, there are at least four areas which require further study. First, MOM should have a more complete range checking subsystem. The current range checking in use is primitive, and does not take into account individual station characteristics, or seasonality. Second, MOM should be expanded by adding buddy checking and time checking methods of validation. These features would allow MOM to be more functional, and help to find more errors. Third, MOM should be delivered out of the test environment and into the working environment. Currently MOM is still running in CLIPS interactively, and testing has been taking place using batch files. A delivery environment for MOM would mean better run time, and a capacity to test more data. Finally, and most ambitiously, an error correction facility could be implemented.