A Meteorological Knowledge Discovery Environment*

Alex G. Büchner[§], J.C.L. Chan[†], S.L. Hung[‡], John G. Hughes[§]

[§] Northern Ireland Knowledge Engineering Laboratory, University of Ulster, UK
[†] Centre for Environmental Science and Technology, City University of Hong Kong, China
[‡] Department of Computer Science, City University of Hong Kong, China

Email: ag.buchner@ulst.ac.uk

Abstract

A meteorological knowledge discovery environment is presented which allows the experimentation with various textual and graphical geophysical data, as well as the incorporation of different types of data mining models. The architecture of the platform contains a data warehouse and a knowledge discovery component as its two major modules. The first has been designed to handle and store meteorological information in a multidimensional materialised view which is created after various statistical and artificial intelligence related techniques have been applied. The latter supports meteorological analyses and prediction of geophysical phenomenon. To show the applicability of the created environment, two meteorologically interesting trial runs are presented, one covering forecasting, one covering nowcasting.

1.1 Introduction

Geophysical data is the most important material meteorologists use to model the behaviour of the earth's atmospheres and oceans. While most research dedicated to explanation and prediction has been based on the application of a specific statistical or artificial intelligence technique, only a few endeavours have tackled the holistic nature of the subject. One possibility of approaching this target is the application of knowledge discovery techniques, or, as P. Storlotz et. al. have put it ([Sto95]): "The important scientific challenge of understanding global climate change is one that clearly requires the application of knowledge discovery and data mining techniques on a massive scale".

Due to the highly heterogeneous nature of the data and the vast amount of available domain expertise, traditional data mining techniques by themselves have proven infeasible. Additionally, due to the large quantity of available historical data, discovery of knowledge from a virtual data view as created in distributed and heterogeneous databases and presented in [Büc96, Büc97, and Cha95] supersedes the capacity of up-to-date algorithms and hardware. An alternative approach, which has proven successful in other disciplines such as finance, retail and manufacturing, is the discovery of knowledge from a materialised view, represented in a data warehouse.

We have followed that approach and designed the Meteorology And DAta Mining Environment (MADAME), which resulted in a promising platform for further research being carried out in this area. The work was motivated by a project the authors were involved in, whose objective was to establish the feasibility of forecasting high intensity rainfall over different areas of the territory of Hong Kong using data mining techniques with a view of improving the existing landslide warning system ([Cha98]). The material of this project is used to show the applicability of the designed and developed environment.

The chapter is outlined as following. First some meteorological background is given, which recapitulates some geophysical phenomena, describes available textual and graphical data sources, and presents related work. Then, in Section 1.3, the MADAME's architecture is described, which contains a data warehousing and knowledge discovery component. These two parts will be described in more detail in Sections 1.4 and 1.5, respectively. To show the applicability of the environment, two

* This work has been partially funded by the Geotechnical Engineering Office, Hong Kong SAR Government, China.

representative trial runs have been carried out in Section 1.6 — one in the area of forecasting (mid to long term prediction) and one in the field of nowcasting (short term prediction). Section 1.7 concludes the chapter summarising the carried out research and outlining further work.

1.2 Some Meteorological Background

The intention of this section is to give some fundamental meteorological background for better understanding of the naturally complex subject. Most of the examples used, stem from the project mentioned above. Although this might seem a limitation of the generic applicability, almost all artefacts have their equivalents in related activities.

Different geographical areas are influenced by different meteorological conditions, which influence local weather scenarios. Depending on the location, weather prediction for a certain region can be of different complexity. Various parts of the entire Asian continent, among others, is influenced by monsoon climate, which can result in very heavy rainfall (up to >100mm per hour) during the summer period. The prediction of that type of conditions is of much shorter time scale (hence nowcasting) than long term predictions during the winter or in more continental regions (hence forecasting). Both techniques can be tackled with various approaches, for example, human expertise, formal models, simulation, persistence (also known as null adaptation), or hybrids thereof.

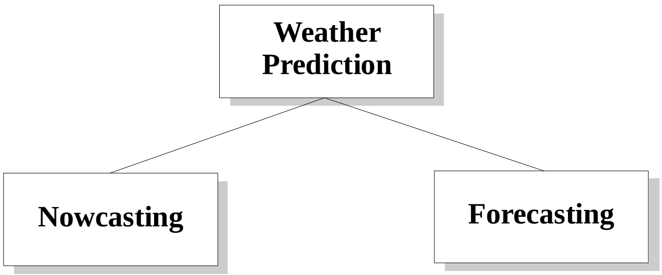

Figure 1.1. Taxonomy of Weather Predictions

Currently most weather forecasting centres use a mixture of human observation of recorded data based on their previous experience and IT supported decision support tools. These two components — data and existing support mechanisms — are described in the following two subsections.

1.2.1 Available Data Sources

The data usually available for predicting weather conditions on whatever time scale is highly diverse and consists of five main sources, which can be sub-divided into textual and graphical data.

Amounts of rainfall are measured at rainfall stations, the density of which can vary quite enormously over the different countries or other observed areas. Automatic weather stations record information on the ground about air and wind related measures, such as wind direction, wind speed, gusts, temperature, wet-bulb temperature, dew point, relative humidity, rainfall, and the mean sea-level pressure. The upper air data provides information about pressure, geopotential height, wind direction, wind speed, temperature, dew point, and relative humidity. The measures, telemetered by radiosondes which are used world-wide to create synoptic weather maps, are provided every 6 hours for the wind-related data and every 12 hours for all other measures.

In addition to these conventional data, every 6 minutes, a 256km range 3km CAPPI radar reflectivity picture is taken and stored as a bitmap file (256 by 256 pixels). Two example radar pictures – one showing little, one showing heavy overcast – are shown in Figure 1.2 below.

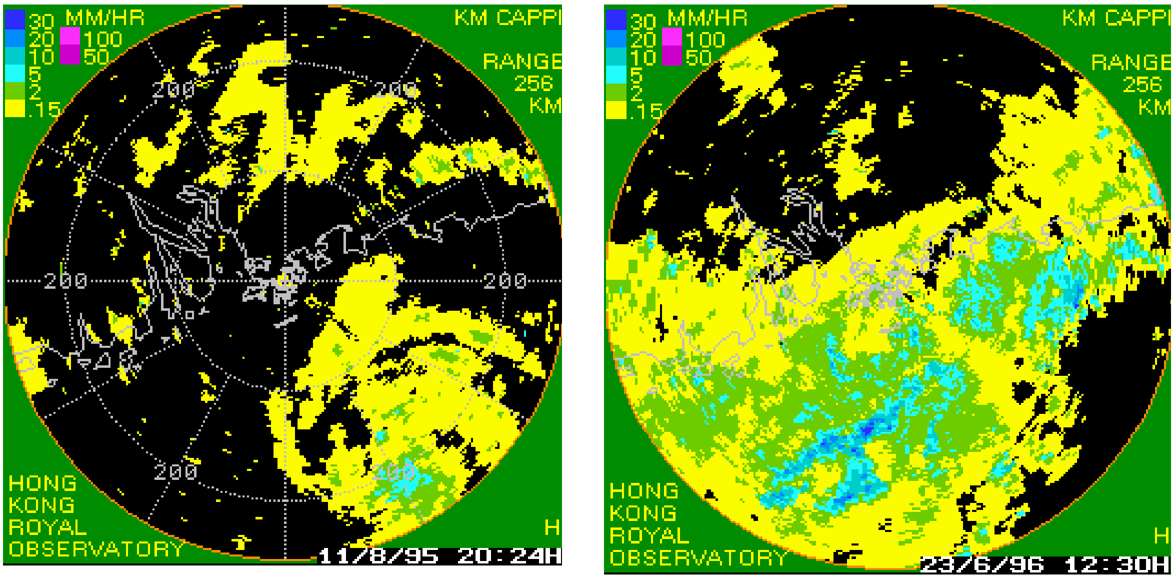

Figure 1.2. CAPPI Radar Reflectivity Pictures

The most useful information which is implicitly stored in radar pictures are the total amount of rainfall and the location(s) of heavy rainfall. The acquisition of that knowledge will be described in more detail at the information extraction stage in Section 1.4.2.

Additionally, satellite imagery at the horizontal resolution (infrared as well as visible as shown in Figure 1.3) are being taken every (6) hour(s). These pictures contain information about water vapour and implicitly about the type of the current cloud structure. The World Meteorological Organization classifies clouds according to their appearance into 10 genera based on the main characteristic forms of clouds (cirrus, cirrocumulus, cirrostratus, etc.). Each of the genera comes in one or more of 14 species, depending on peculiarities in the shapes or internal structures of the clouds ([Bat84]).

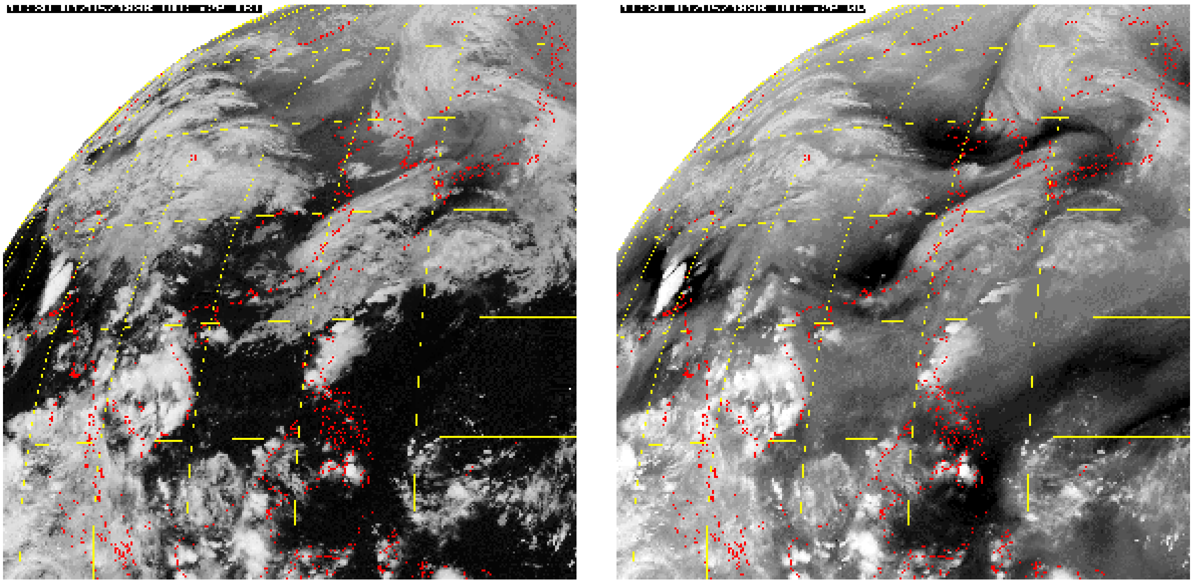

Figure 1.3. Infrared Satellite Pictures

The five available data sources contain a wide range of schematic and semantic heterogeneities, which have to be resolved before the data can be used as input for any data mining exercise. The main types of heterogeneities are incompatible units (measures and timezones) as well as spatial inconsistencies. The latter does not only affect x and y co-ordinates, but adds a third dimension, viz. (geopotential) height. The second crucial aspect of the data is that of temporal granularity. While some measures are taken in 5-minute intervals, others are recorded only twice a day. Finally, the amount of missing values within the data needs special attention, while noise rarely exists in meteorological data (see also Section 1.4.3).

A summary of the diversity of the available data from Hong Kong Observatory — after resolving schematic naming conflicts — is given in Table 1.1. As can be seen from the table below, the graphical data in the project was partly incomplete, which was caused by extraction problems from the raw data at the Observatory. Only data from disastrous events was made available for feasibility purposes.

Type	Source	From	To	Interval	Total number of records	Field Name	Field Unit
Textual	*Rainfall*	04/84	09/96	5 min	4,110,912	DateTime	HKT
						Station	1..75
						Value	1/10 mm
	AWS	04/87	09/96	60 min	263,544	DateTime	HKT
						Station	1..24
						WindDirection	10º steps
						WindSpeed	1/10 m/sec
						Gust	1/10 m/sec
						Temperature	1/10 ºC
						WetBulbTemp	1/10 ºC
						DewPoint	1/10 ºC
						RelativeHumidity	%
						Rainfall	1/10 mm
						MeanSeaLevelPressure	1/10 hPa
	Sonde	04/87	09/96	720 min	21,962	DateTime	GMT
						Pressure	hPa
						GeoPotentialHeight	m
						Temperature	1/10 ºC
						DewPoint	1/10 ºC
						RelativeHumidity	%
	Wind	04/87	09/96	360min	43,924	DateTime	GMT
						Pressure	hPa
						WindDirection	10º steps
						WindSpeed	1/10 m/sec
Graphical	*Radar*			6 min		AmountOfRainfall	mm
						Entropy	0..1
						Direction	10º steps
	Sattelite			360 min		CloudType	<set of cloud types>
						WaterVapour	g/kg

Table 1.1. Available Data Sources

1.2.2 Related Work

Analytical weather forecasting is based on solving extremely complex dynamical mathematical models, which demand substantial computing power, detailed atmospheric measurements, and accurate updates of various boundary conditions ([Cho97]). Although very few weather forecasting centres actually facilitate artificial intelligence approaches to perform forecasting, there has been some substantial work done in that area. A representative set of approaches is described, which has tackled meteorological problems in the past[1]. The endeavours are sub-divided into neural networks, case-based reasoning, and numerical models.

[1] In addition to the briefly described systems here, there are various commercial systems available (for instance Merlin or Storm), which promise the predictability of weather-related scenarios and which are supposedly based on some artificial intelligence techniques. Since the objective of these systems are of monetary nature, no detailed information is available of what exact techniques are being facilitated.

1.2.2.1 Neural Network Approaches

[Atl90] have used meteorological data to test their sophisticated artificial neural networks to forecast weather scenarios. In their study, the neural network(s) also performs non-linear regression among load and weather patterns. When compared with classification methods such as the classification and autoregression trees, the network shows a superior performance in terms of accuracy. [McC92] have used a similar technique to forecast thunderstorms.

The Tampere University of Technology has carried out a study which endeavoured to model the short term district heat load forecasting, for which they applied multi-layer perceptron networks ([Leh94]). The objective was to create a system which can be built with a reasonable amount of example data. The different factors affecting the structure and construction of the model are discussed. The model proves to be working well when tested with independent test data.

[Cho97] has used a recurrent sigma-pi neural network to built a prototypical nowcasting system. The input nodes represented 28 values derived from radar images (see also Section 1.4.2) as well as three averaged rain gauge values. The network is based on back-propagation learning and achieved reasonably good results.

The Hong Kong Polytechnic University is currently carrying out a study, which has the following objectives to explore possible applications of multiple strategy machine learning and discovery methods including: (1) methods of knowledge-based approximation and automatic function construction under the control of domain expert knowledge and heuristic rules, (2) ideas of variable generation and variable reduction by the aid of an exploration matrix and instance tensor to facilitate discovery of new forecasting rules, (3) performance improvement approaches which include self-error-correction algorithms and rule improvement algorithms for the refinement of discovered rules, (4) hybrid approaches of existing algorithms and exploring methods, such as integrating the approaches of the inductive paradigm and the connectionist paradigm. The approach uses mainly advanced statistical techniques and the connectionist paradigm ([Liu96]).

1.2.2.2 Case-base Reasoning Approaches

[Jon95] have built a workbench which serves as an intelligent assistant for meteorologists: “a kind of memory amplifier that allows meteorologists to locate and analyze historical situations”. The system is fully based on case-based reasoning techniques. The system handles 2,500 cases over 3½ years and another 10,000 are currently being incorporated. The workbench provides facilities to run (manual or automatic) queries against the case-base and to return the scenario(s) with the closest match(es).

1.2.2.3 Numerical Approaches

The vast majority of built prediction models which are used in weather forecasting centres are based on numerical models, for example, [Yeu89]’s simulation model which is used in Hong Kong. [Col89] have built a complex numerical prediction model using satellite images and radar information; [Lam84] have used radar pictures only; and [Rod87] have used their model to describe characteristics of rain storms using temporal and altitudinal data as input.

1.3 MADAME's Architecture

To incorporate available temporal, spatial and altitudinal data as described in Section 1.2.1 and support machine learning algorithms as outlined in Section 1.2.2, a meteorological knowledge discovery architecture has been designed, which is depicted in Figure 1.4.

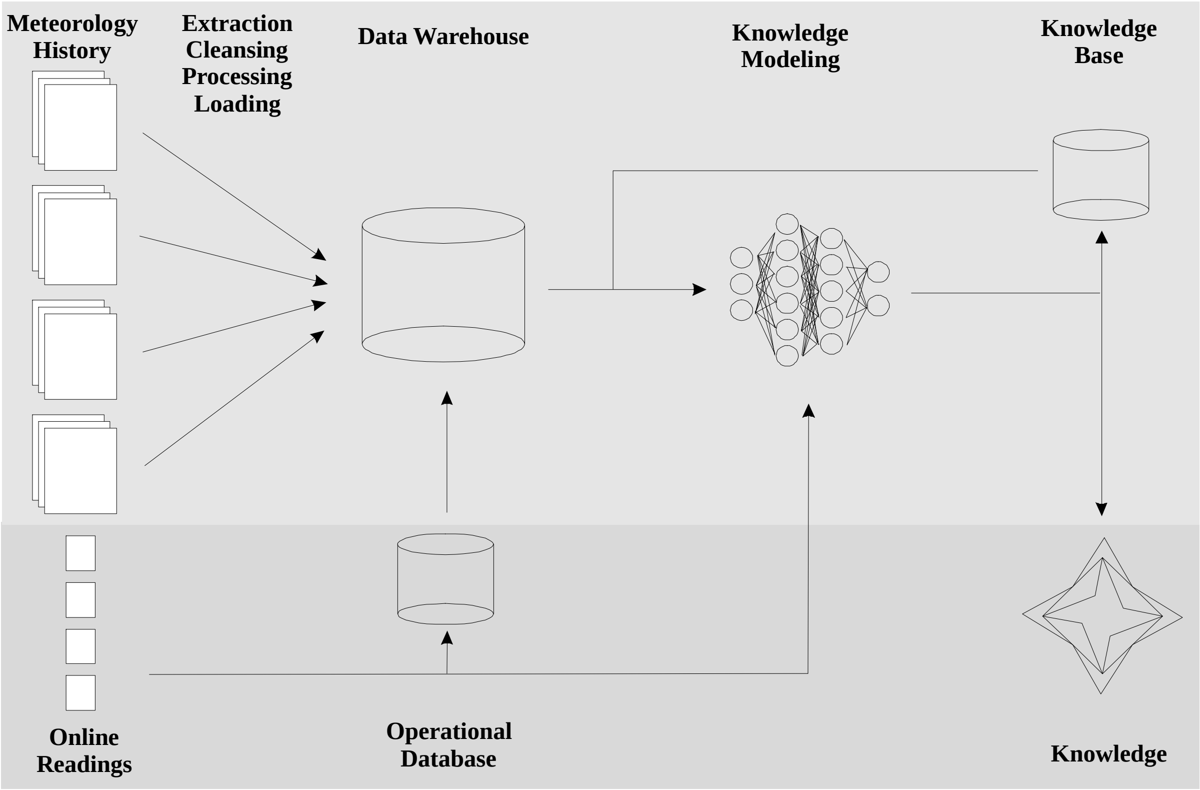

Figure 1.4. A Meteorological Knowledge Discovery Architecture

The top part of the architecture deals with historical meteorological information, which consists of a battery of data sources as described above. Due to their extreme heterogeneity and complexity, various extraction, cleansing, processing, and loading operations have to be performed. These transformations are described in more detail in Section 1.4. The type of transformations depends heavily on the design of the data warehouse, which has to support multiple materialised views (see Section 1.4.1). The remaining parts — knowledge modelling and the knowledge base — form the knowledge discovery component, which is described in Section 1.5.

The bottom part of the architecture represents the online component. The online readings are transformed identically to the historical data. The data warehouse is updated on a regular basis (for example, after every monsoon season or a shorter interval if necessary). The readings are then run through the appropriate knowledge model, whose output is used for analyses and predictions, and which is, if relevant, stored in the knowledge base.

This environment can either be connected to any existing meteorological prediction system and provide complementary information on forecasting or can be used as an experimental stand-alone system. Further parts of this architecture are visualisation tools and other reporting and / or analysing techniques.

1.4 Building a Meteorological Data Warehouse

In this section, individual parts and transformation operations of the data warehousing components are described. For illustration purposes, examples from the project carried out in Hong Kong are used as mentioned in the introduction. The arsenal of pre-processing tools has proved sufficient to perform standard meteorological exercises; for more specialised tasks, enhancements might be necessary, which can easily integrated.

1.4.1 The Design

Although the design of a meteorological data warehouse depends crucially on the purpose of the environment, most parts are of a rather generic nature. The most typical dimensions of geophysical data are time, spatiality and pressure[2]. An extensible hypercube which shows these three dimensions is depicted in Figure 1.5. Each dimension contains a number of attributes, as well as factual summarisation information.

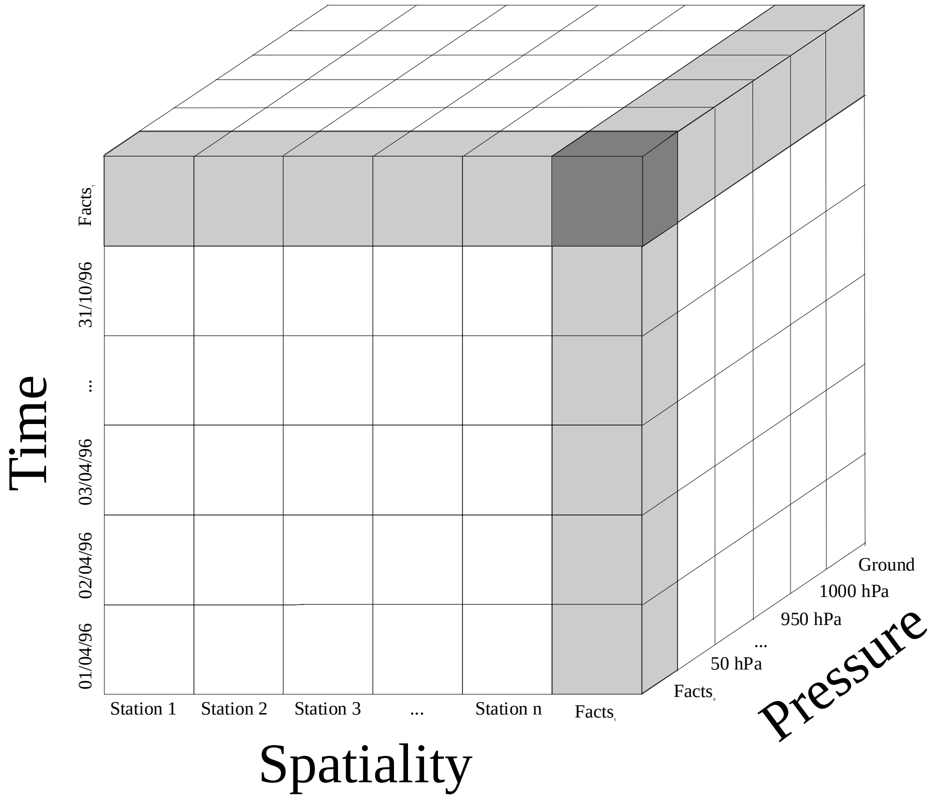

Figure 1.5. A meteorological Hypercube

Internally, each dimension is represented as a relation, which is connected to a fact table. This star schema is sufficient for one given set of scenarios, which uses data input of the same granularity ([Ber97]). For more advanced operations, as necessary in the meteorological context, a snowflake schema is required, which supports multiple granularities. Snowflake schemas provide a refinement of star schemas where the dimensional hierarchy is explicitly represented by normalising dimension tables ([Cha97]). For instance, in the case of the Hong Kong data, stations can be located in various spatial schemes ([Cho97]) or in a Cartesian co-ordinate system and time can be grouped into hourly intervals, 6 hour ranges or days, among others. Since the modelling of the schema itself is a database-related problem, further details are omitted here.

1.4.2 Information Extracting

The purpose of information extraction is the derivation of new values from existing data ([Ana98]). An example is the calculation of the adiabatic lapse rate derived from two temperatures t_1 and t_2 at two heights h_1 and h_2 as $alr = -\frac{t_1 - t_2}{h_2 - h_1}$, which indicates temperature inversion. More interesting though, is the information being possible to extract from graphical material.

As mentioned at the data prospecting stage in Section 1.2.1, the most important meteorological information being possible to extract from radar pictures is the amount of rainfall, the position of the rainfall area with the strongest impact and the cohesion of rainfall areas. After converting the radar information into Cartesian co-ordinates, the amount of rainfall is calculated as the sum of all rainfall pixels, in mm per hour. Cluster detection ([Miy90, Mes94]) can then be used to find all rainfall areas. The order of their impact (represented as the weight w_i) is computed as the ratio of intensity and size.

[2] Meteorologists measure altitude at a certain pressure and not vice versa. Exceptions are measures taken at ground stations.

The direction is the relative position of the centre of gravity of the observed area to the centre of attention *m*. The cohesion of one radar scenario is calculated as follows

$$c(x,m) = \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta(x_i,m)}{w_i}$$

where *n* represents the number of rainfall areas, and $\Delta(x_i, m)$ the distance from the centre of gravity of each rainfall area to the centre of attention *m*, which was in the project's case Hong Kong.

[Cho97] has applied a similar technique, which has been adapted from [Lam84]. It converts the radar echoes into rainfall rates with greater density in the immediate vicinity of the centre of attention *m*, with less dense blocks further away from *m*. In the case of the Hong Kong data, this led to 28 new values which were used as input for a neural network, as described in Section 1.2.2.1.

Similarly, water vapour measures can be extracted from satellite pictures. Also, a supervised artificial neural network can be applied to classify cloud genera. This particular extractor has not been implemented yet, but is part of further work (see Section 1.7).

1.4.3 Data Cleansing

Data cleansing is concerned with the treatment of incorrect and irrelevant measures. Outliers (for example, negative temperatures in a subtropical climate) and errors (for instance wind directions greater than 360^o or relative humidity greater than 100%) are the most often found faulty values.

Irrelevant measures are more dependent on the geographical area the meteorological knowledge discovery environment facilitates than errors. For instance, since major rainfall-related catastrophes in Southeast Asia happen between April and September, only those months have been chosen for further evaluation. Further, before April 1987 only rainfall data was available, which is — standalone — rather fruitless for data mining purposes, and thus has been ignored. This resulted in a reduction of the original data of more than 50%.

Another type of incorrect measurement is noise. Although more feasible in other domains in which equipment with less accuracy (including human beings!) is used, it can have some impact in the field of meteorology. For example, weather radar detects rain in the atmosphere and determines the amount and distance from the time delay and signal strength of microwave echoes. The radar system has to be manually re-calibrated every two to three years, which incorporates some mismatch in the data when compared with previous measures. This type of modification has to be considered in the data cleansing module.

NULL values have to be dealt with, for example, the rain gauge data as well as other measures may often be missing for one or two time slots, especially in bad weather conditions. These values can either be replaced with default values, considered in the decision making mechanism or filled in with a calculated value. The latter is usually based on statistical methods such as interpolation, which is described in more detail in the following section.

1.4.4 Data Processing

The two main objectives of data processing are resolving semantic and schematic heterogeneities in the data and mapping the data onto different materialised views.

Semantic interoperability among different data sources is guaranteed by applying standard conversion functions. For instance, for all date and time values the chosen canonical form is usually the local time zone (radar information is stored in GMT), all pressures are given in hPa, all temperatures in degree Celsius, and so forth.

To resolve the spatial heterogeneity between the different weather station types (rainfall stations and automatic weather stations), a lookup table has been created which is based on the co-ordinate system of area of observation and divides the covered area into an appropriate number of quadrants. For the carried out project the Hong Kong territory was organised in a 7 by 5 matrix. [Cha97] suggested three further schemes based on more meteorological grounds using two, three and ten areas, respectively. These different schemes are then modelled and stored in the built data warehouse.

Altitudinal data, that is tuples distributed over different heights (pressures), has to be stored as such. But, depending on the knowledge model built, different ranges are of interest. For example, the involved domain expert in Hong Kong was particularly interested in the data at a pressure of 850 hPa, which gives a representational height for rainfall characteristics. [Che93] has carried out a study, which concentrated on the 500 hPa height field.

To overcome the temporal heterogeneity across data sources, various interpolation and summarisation steps have to be performed. Depending on the interval length, time-related data has to be either summarised or interpolated. To perform the discovery of general characteristics (associations and classifications) of heavy rainfall, a 24 hour interval has been used, which requires summarising and aggregating of data. To perform the discovery of sequential patterns, an hourly interval has been used, which, of course can be extended or shrunk. The data with a granularity of less than 60 minutes had been summarised and aggregated analogous to the 24 hour interval; data with a greater interval had to be interpolated. For all non-radial measures the standard interpolation has been used; for radial data (wind directions) the value was distributed over missing values to be filled in. Again these are temporal intervals used for the project carried out; every other combination of granularities which is needed as data mining algorithm input is supported by the data warehouse.

1.4.5 Data Loading and Refreshing

Finally, after extracting, cleansing and transforming, data must be loaded into the data warehouse. Additional processing, such as checking integrity constraints, sorting, indexing, etc. may be required. The mainly databases-related details are omitted and can be found in [Cha97].

The refreshing of the data warehouse has to be performed at regular intervals to re-train the built knowledge model(s). Typical intervals are daily, weekly, monthly, or after every weather season. An alternative approach, which is feasible in meteorological contexts, is the refreshing after every important event, for example, landslides caused by heavy rainfall or a high-scale storm warning.

The loaded data warehouse forms a smooth interface to the knowledge discovery components, which are described in the following section.

1.5 The Knowledge Discovery Components

1.5.1 Knowledge Modelling

Due to the openness of the architecture there is no limit to the type and number of supported knowledge modelling techniques. The only pre-requisite is the acceptance of data input from the materialised view and output which is actionable in meteorological prediction terms. A further optional requirement is the incorporation of domain knowledge as described in Section 1.5.2. A set of data mining algorithms used in the carried out project is described in the sequel.

Two general data mining exercises have been performed. The first was looking for patterns at 24-hour intervals. This discovered knowledge is useful for forecasting purposes; it gave a first impression of the data being mined, and also gave some insights into general characteristic of heavy rainfall. The second was dealing with the more interesting 1-hour intervals that also considered the temporal dimension of the data, which is relevant for nowcasting. Both data mining runs considered data of the 850 hPa height field. In order to validate the discovered patterns (see Section 1.6) the data has been split up into two parts: the training data encompasses the years 1990 to 1996, the testing data encompasses 1987-1989. The reason for this particular division is based on the reverse chronological way the data was provided by the observatory.

1.5.1.1 Modelling Forecasting Knowledge

The pre-processed data has been clustered into the three groups ‘low’ (less than 20mm for the entire observed region per day), ‘moderate’ (between 20 and 100mm), and ‘heavy’ (more than 100mm) rainfall. Due to the fact that the amount of ‘heavy’ rainfall cases exceeds the number of the other cases, these cases were artificially balanced to a quasi-equal level.

To discover associations across the pre-processed data, the general rule induction algorithm GRI ([Dom96]) has been applied. For the purpose of this data mining exercise the declared rainfall groups have been chosen as the antecedent. An example rule is

```
RainGroup == heavy
      Sonde_Temperature > 11,35°C and
      AWS_RelativeHumidity > 88,369% and
      AWS_MeanSeaLevelPressure < 1009,55 hPa
      (support:15,85%  confidence:76.0%)
```

The pre-underscore part of each attribute indicates the source (Sonde, AWS, Wind, and Rainfall), while the post-underscore part is the field itself. The support value represents the ratio of the number of the records in the database for which the rule is true to the total number of records in the database. Confidence expresses the belief in the consequent being true for a record once the antecedent is known to be true. With a minimal support threshold of 1% and a confidence threshold of 50%, 47 rules were found by GRI.

To discover classifications in the data, three different techniques were applied, all being based on the C5.0 classification algorithm ([Rul98]). First, classification trees were used and a set of rules derived. The strongest rule being discovered is shown below. The notation of accuracy values and field names is identical to that described for associations.

```
Rule #4 for heavy:
      if   AWS_WetBulbTemp <= 26,115%
      and  AWS_RelativeHumidity > 85,514%
      and  Sonde_Temperature > 17,95°C
      and  Wind_WindSpeed > 0,54 m/sec
      then -> heavy
      (support:28,96%  confidence:89,1%)
```

The second approach was to apply a back propagation neural network ([Big96]), which led to a model with an input layer with 25 neurons, a hidden layer with 7 nodes, and output layer with 3 neurons. Each input node represents one field in the materialised data view with a certain weight allotted to it; each output node represents a value of the classification label (heavy, moderate, or low). The training of the artificial neural network was stopped after the lowest error was found in the training set and the network had not improved for persistent cycles.

The last approach used the discovered neural network as input for the classifier which slightly improved the accuracy of the C5.0 model. Comparing the accuracies of the three approaches on the forecasting training data gives the following result, which shows a clear advantage of the rule-based approaches over the neural network (see Figure 1.6). Due to the fact that the output field of all three models is symbolic, the accuracy represents the total number of correct cases expressed as a percentage of the total number of cases.

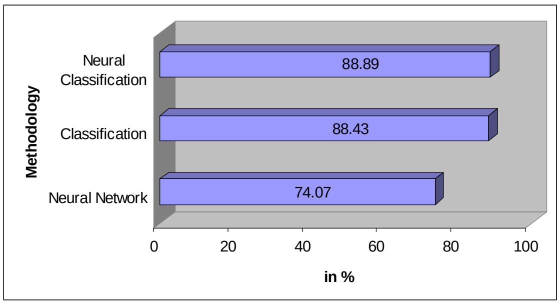

Figure 1.6. Accuracy Comparison of different Forecasting Models

1.5.1.2 Modelling Nowcasting Knowledge

The main differences between mining the 1-hour data and the 24-hour data are the increase in input tuples by a factor 24 and its temporal nature. An additional 'hour' field represents this time component. To exploit this time field for predicting purposes a history operation has been performed, which puts previous values of one or more fields into the current record and adds new fields to each record that passes through it. To find patterns which consider information about up to 4 hours before heavy rainfall occurs the offset (which indicated the latest record prior to the current record from which field values should be extracted) of 4 and span (which indicates the number of records from which to extract values) of 1 has been used.

Again, three types of data mining exercises to discover sequential patterns which were then used to classify the data into 'low', 'moderate', and 'heavy' rainfall have been performed. A back-propagation neural network, the C5.0 classifier, and the synergy of both approaches, that is, the output of the neural network have been used as input for the classifier. Similarly to the artificial neural network used above, the network used all data source input fields (considering the temporal nature this led to 78 input neurons), 37 neurons in the hidden layer and three output nodes[3].

The discovered patterns have the same format as the ones in the previous section. The only amendment is the format of temporal components. Every variable without an extension belongs to the current time window; every variable with an extension of the format <variable>__*x* (1 ≤ *x* ≤ 4) belongs to the time window *x* hours ago. An example rule is as follows.

```
Rule #17 for heavy:
    if   AWS_Dewpoint <= 25.025°C
    and  AWS_MeanSeaLevelPressure <= 10106.5 hPa
    and  Wind_WindDirection == SouthEast
    and  AWS_WindSpeed__2 <= 6 m/sec
    and  AWS_RelativeHumidity__1 <= 85.571%
    and  AWS_Rainfall__3 <= 11.833mm
    and  Rainfall__2 > 2mm
    and  Rainfall__2 <= 139mm
  then -> heavy
  (support: 10,38%  confidence:95,2%)
```

Depending on the support and confidence thresholds used, the number of discovered rules varied, but did not exceed the number of rules discovered in the forecasting scenario. The comparison of the accuracies of the three models on the nowcasting training data is shown in the figure below, which shows a clear advantage of the classification approach over the neural network. It also shows that it is possible to improve the quality of the accuracy of the classifier using the generated artificial neural network.

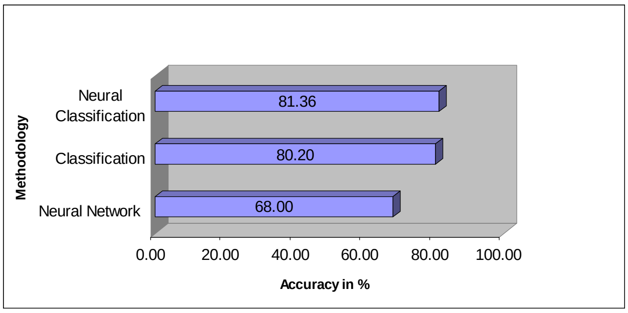

Figure 1.7. Accuracy Comparison of different Nowcasting Models

[3] The system used, which is based on an expert system containing rules for dynamic configuration, generates the topology of the artificial neural network automatically.

1.5.2 Domain Knowledge

Domain knowledge elicitation and incorporation is an essential step in every data mining process. The major objectives of this phase are reducing the dimensionality of the search space and constraining the rule space, and thus improving the knowledge quantity and quality. The collected domain knowledge is then stored in the knowledge base.

In general, available domain knowledge can be divided into extrinsic (subjective) and intrinsic (objective) knowledge ([Sil96]). In the domain of meteorology, the first is relatively easy to capture and harness, whereas the latter needs some more attention. Various samples of explicit knowledge have been used implicitly at the data pre-processing stage as described in Section 1.4. Examples are often well known facts (for the involved experts), such as that there is no heavy rainfall between October and March in sub-tropical areas in the northern hemisphere, that some weather stations are only used for testing purposes, etc. An example of extrinsic domain knowledge is given in the figure below, in which numeric wind directions are clustered into 8 wind direction groups using the MKS domain knowledge format[4].

```
WindDirection User_Interval
START
    gt   0.0 le  22.5 N
    gt  22.5 le  67.5 NE
    gt  67.5 le 112.5 E
    gt 112.5 le 157.5 SE
    gt 157.5 le 202.5 S
    gt 202.5 le 247.5 SW
    gt 247.5 le 292.5 W
    gt 292.5 le 337.5 NW
    gt 337.5 le 360.0 N
END
```

Figure 1.8. Extrinsic Domain Knowledge Example

Much harder is the incorporation of intrinsic knowledge, that is knowledge a meteorologist uses on a day to day basis, because it has to be put down in a formalised way. There are various different mutually inclusive possibilities concerning how to elicit a domain expert's knowledge in the data mining process, for example, attribute constraints, hierarchies, or thresholds ([Ana95]). For the purpose of this study, which aims to show the feasibility of the application of data mining in the meteorology domain, only extrinsic knowledge has been incorporated. Further intrinsic knowledge will be elicited at the refinement stage.

1.6 Prediction Trial Runs

To show the applicability of the built models, it was concentrated on the 1-hour interval nowcasting data, since that provides the most helpful information to improve the prediction of heavy rainfall in monsoon areas like Hong Kong. In order to perform this trial, two different approaches were run. First, we use the data of 1987 to 1989 to test the built models, which have been built with the training data from 1990 to 1996. Second, we pick a set of days when landslides occurred and match the values against the discovered rules.

1.6.1 Nowcasting of Heavy Rainfall

To test the built nowcasting models the 1987 to 1989 data was pre-processed in exactly the same way as the training data. This pre-processed data was then used as input for the knowledge models developed during the data mining exercise and accuracies on the testing data compared to those of the training data. The outcome of the models are listed in the following table:

[4] MKS is the Mining Kernel System, developed at the authors' laboratory ([Ana97])

Methodology	Accuracy Training Data	Accuracy Test Data	Accuracy Test Data (high only)
Neural Network	68.00 %	66.53%	68.75%
Classification	80.20 %	50.00%	74.71%
Neural Classification	81.36 %	54.44%	53.57%

Table 1.2. Accuracy of Testing Data Set

These results can be interpreted as following. The neural network performs better when applied on the entire data set, whereas the classifier outperforms the net, when only looking for high rainfall, which was the objective of the study. The full list of analyses of the testing of the three models is given in the final project report ([Cha98]).

1.6.2 Landslide Nowcasting

To simulate a trial forecast, the data from the 12th of August 1995 has been used, when 60 landslides occurred, 2 of which were most disastrous. The three models were run for data of every single hour and the outcome has been monitored in the following table. The official flooding warning was raised on 12th of August at 04:45.

Day / Time	Neural Network	Classifier	Neural Classifier
11/04 17:00	High	Low	Low
11/04 18:00	Low	Low	Low
11/04 19:00	Low	Low	Low
11/04 20:00	Low	Low	Low
11/04 21:00	Low	Low	Low
11/04 22:00	Low	Low	Low
11/04 23:00	Low	Low	Low
12/04 00:00	Missing	Missing	Missing
12/04 01:00	Low	Moderate	Moderate
12/04 02:00	High	High	High
12/04 03:00	High	High	High
12/04 04:00	High	High	High
12/04 05:00	High	High	High

Table 1.3. Nowcasting Trial Run Results

The outcome of the test run can be interpreted as following. All three models triggered the alarm at the same time, which is not surprising, since one of the most destructive catastrophes during summer 1995 has been chosen. The official alarm was raised shortly after the first landslides occurred, hence all three models triggered the alarm too early, which can be caused by the following not mutually exclusive factors:

- the models being built are over-sensitive and triggered at a threshold, which is too low.
- more than one consecutive trigger is required to call a serious alarm.
- the missing values at midnight are highly likely to have disrupted the findings of sequences.

No matter what the exact cause for the behaviour is, a major fact can be drawn from the observation: The outcomes clearly show the feasibility of the applicability of the designed and developed meteorological knowledge discovery environment.

For the given time, the three weather stations at which the heaviest rainfall occurred between 01:00 and 02:00 were looked up. The amount of rainfall is listed in the Table 1.4. In fact, one of the landslides occurred in the area of the New Territories in which the stations N12 and N14 are based.

Station	Rainfall between 01:00-02:00
N14	115 mm
R42	100 mm
N12	100 mm

Table 1.4. Location of Heavy Rainfall

1.7 Conclusions and Further Work

A meteorological knowledge discovery environment has been developed, which includes a data warehousing and a data mining component. Both parts have been described in detail and the applicability of the system has been shown. The architecture has been designed in a way that it is highly extensible in order to allow performing as many meteorological experiments as possible. This holds for the data warehousing side as well as the machine learning algorithms and the knowledge base. The case study has also shown some advantages of hybrid data mining, which combines various artificial intelligence techniques at different levels.

Further work will be based on the existing architecture. Work in progress includes the information extraction from satellite images to detect cloud genera as outlined in Section 1.4.2. Also, the domain knowledge incorporation has to be improved which will be performed in collaboration with involved meteorologists. In addition to the two performed case studies a third type has to be carried out to show the applicability of the environment. Seasonal prediction involves the discovery of general patterns for a certain time of the year in a certain area.

Acknowledgements
We would like to thank the Geotechnical Engineering Office at Hong Kong SAR Government for funding the work carried out. We would like to acknowledge the support from the Royal Observatory Hong Kong for providing us with large quantities of meteorological data. We would also like to thank Hughs Man from Global Systems Company, Hong Kong for his support in pre-processing some of the vast amounts of that data.

1.8 References

[Ana95] S.S. Anand, D.A. Bell, J.G. Hughes. The Role of Domain Knowledge in Data Mining, in *Proc. of the 4th Int'l. ACM Conf. on Information and Knowledge Management*, pages 37-43, 1995.

[Ana97] S.S. Anand, B.W. Scotney, M.G. Tan, S.I. McClean, D.A. Bell, J.G. Hughes, I.C. Magill. Designing a Kernel for Data Mining, in *IEEE Expert Intelligent Systems & their Applications*, 12(2):65-74, 1997.

[Ana98] S.S. Anand, A.G. Büchner. Decision Support Through Data Mining, Financial Times Pitman Publisher, 1998.

[Atl90] L.E. Atlas, R. Cole, Y. Muthusamy, A. Lippman, G. Connor, D.C. Park, M. El-Sharkawi, R.J. Marks II. A performance comparison of trained multi-layer perceptrons and classification trees, in *Proc. of the IEEE*, 78:1614-1619, 1990.

[Bat84] L.J. Battan. Fundamentals of Meteorology, 2nd Edition, Prentice-Hall International, 1984.

[Ber97] A. Berson, S.J. Smith. Data Warehousing, Data Mining and OLAP, McGraw Hill, 1997.

[Big96] J.P. Bigus. Data Mining with Neural Networks, McGraw-Hill, 1996.

[Büc96] A.G. Büchner, S.S. Anand, D.A. Bell, J.G. Hughes. A Framework for Discovering Knowledge from Distributed and Heterogeneous Databases, in *Proc. IEE Coll. on Knowledge Discovery and Data Mining*, London, pages 8/1-8/4, 1996.

[Büc97] A.G. Büchner, B. Yang, S. Ram, D.A. Bell, J.G. Hughes. A Holistic Architecture for Knowledge Discovery in Multi-Database Environments, in *Proc. ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'97)*, Tucson, AZ, page 87, 1997.

[Cha95] P. Chan, S. Stolfo. Learning arbiter and combiner trees from partitioned data for scaling machine learning, in *Proc. 1st Int'l. Conf. on Knowledge Discovery and Data Mining*, pages 39-44, 1995.

[Cha97] S. Chaudhuri, U. Dayal. An Overview of Data Warehousing and OLAP Technology, Technical Report MSR-TR-97-14, Microsoft Research, 1997.

[Cha98] J.C.L. Chan, S.L. Hung, A.G. Büchner. Feasibility Study on Improved Forecasting of High Intensity Rainfall for Landslip Warning, Final Project Report, City University of Hong Kong, 1998.

[Che93] X. Cheng, J.M. Wallace. Cluster Analysis of the Northern Hemisphere Wintertime 500-hPa Height Field: Spatial Patterns, in *Journal of Atmospheric Sciences*, 50:2674-2696, 1993.

[Cho97] T.W.S. Chow, S.Y. Cho. A Novel Neural Based Rainfall Nowcasting System in Hong Kong, in *Journal of Intelligent Systems*, 7(3-4):245-264, 1997.

[Col89] C.G. Collier, D.M. Goddard, B.J. Conway. Real-time analysis of prediction using satellite imagery, ground-based radars conventioanl observations and numerical model output, in *Meteorology Magazine*, 118(1398):1-8, 1989.

[Dom96] P. Domingos. Using Partitioning to Speed Up Specific-to-General Rule Induction, in *Proc. AAAI Workshop on Integrating Multiple Learned Models for Improving and Scaling Machine Learning Algorithms*, pages 29-34, 1996.

[Jon95] E.K. Jones, A. Roydhouse. Intelligent Retrieval of Archived Meteorological Data, in *IEEE Expert Intelligent Systems & their Applications*, 10(6):50-57, 1995.

[Lam84] C.Y. Lam. Digital Radar Data as an aid in nowcasting in Hong Kong, in *Proc. Nowcasting-II Symp.*, Norrloping, Sweden, 1984.

[Leh94] O. Lehtoranta, J. Seppala, H. Koivisto, H. Koivo. Neural Network Based District Heat Load Forecasting, Technical Report Tampere University of Technology, 1994.

[Liu96] J. Liu, L. Wong. A case study for Hong Kong weather forecasting, in *Proc. Int'l. Conf. on Neural Information Processing*, pages 787-792, September 24-27, 1996.

[McC92] D.W. McCann. Forecasting techniques, a neural network short-term forecast of significant thunderstorms, in *Weather & Forecasting*, 7(3), 1992.

[Mes94] E. Mesrobian, R.R. Muntz, J.R. Santos, E.C. Shek, C.R. Mechoso, J.D. Farrara, P. Storlorz. Extracting Spatio-Temporal Patterns from Geoscience Datasets, in *IEEE Workshop on Visualization and Machine Vision*, 1994.

[Miy90] S. Miyamoto. Fuzzy Sets in Information Retrieval and Cluster Analysis, Kluwer Academic Publications, Dordrecht, Boston, London, 1990.

[Rod87] I. Rodriguez, P.S. Eagleson. Mathematical models of rain storm events in space and time, in *Water Resource*, 23(1):181-190, 1987.

[Rul98] Rulequest. http://www.rulequest.com, 1998.

[Sil96] A. Silverschatz, A. Tuzhilin. What Makes Patterns Interesting in Knowledge Discovery Systems, in *IEEE Transactions on Knowledge and Data Engineering*, 8(6)970-974, 1996.

[Sto95] P. Stolorz, E. Mesrobian, R.R. Muntz, E.C. Shek, J.R. Santos, J. Yi, K. Ng, S.-Y. Chien, H. Nakamura, C.R. Mechoso, J.D. Farrara. Fast Spatio-Temporal Data Mining of Large Geophysical Datasets, in *Proc. 1st Int'l. Conf on Knowledge Discovery and Data Mining*, pages 300-305, 1995.

[Yeu89] K.K. Yeung, W.L. Chang. Numerical simulation of mesoscale meteorological phenomena in Hong Kong, in *Proc. Int'l. Conf. on East Asia and Western Pacific Meteorology and Climate*, pages 451-460, 1989.