# THE ENVIRONMENTAL SCENARIO GENERATOR (ESG): A DISTRIBUTED ENVIRONMENTAL DATA ARCHIVE ANALYSIS TOOL

***E.A. Kihn[1]*, M. Zhizhin[2] , R. Siquig[3] and R. Redmon[4]***

**1 NOAA/NGDC 325 Broadway E/GC2 Boulder, CO 80305, USA*
Eric.A.Kihn@noaa.gov

*2 RAS/CGDS 3 Molodezhnaya Str, Moscow, 117964, Russia*
jjn@wdcb.ru

*3 Naval Research Laboratory, 7 Grace Hopper Ave, Monterey, CA 93943, USA*
siquig@nrlmry.navy.mil

*4 NOAA/CIRES 216 UCB Boulder, CO 80309, USA*
Rob.Redmon@noaa.gov

## ABSTRACT

*The Environmental Scenario Generator (ESG) is a network distributed software system designed to allow a user to interact with archives of environmental data for the purpose of scenario extraction, data analysis and integration with existing models that require environmental input. The ESG uses fuzzy-logic based search tools to allow a user to look for specific environmental scenarios in vast archives by specifying the search in human linguistic terms. For example, the user can specify a scenario such as a "cloud free week" or "high winds and low pressure" and then search relevant archives available across the network to get a list of matching events. The ESG hooks to existing archives of data by providing a simple communication framework and an efficient data model for exchanging data. Once data has been delivered by the distributed archives in the ESG data model, it can easily be accessed by the visualization, integration and analysis components to meet specific user requests. The ESG implementation provides a framework which can be taken as a pattern applicable to other distributed archive systems.*

**Keywords:** Fuzzy Logic, Data Archives, Data Warehousing, Data Mining, Systems Architecture, Environmental Informatics, Distributed Parallel Computing, Web Services

## 1 INTRODUCTION

Environmental Informatics (Hilty, Page, Radermacher, & Riekert, 1995) is a rapidly expanding area of computer and natural science. The increasing data volumes from today's collection systems and the needs of the scientific community which require the inclusion of an integrated and authoritative representation of the natural environment in their simulations require a new approach to data management and access. Here the natural environment includes elements from multiple domains such as space, oceans, terrestrial weather and terrain. The capability exists today to model a highly realistic environment on a wide range of scales. Systems such as the Master Environmental Library (MEL) (DMSO, 2003), National Virtual Data System (NVDS)( NOAA Satellite & Information, n.d.), and others provide the ability to search for the environmental data sets distributed across the network, but the ability to search for specific "scenarios" (sets of conditions within the archived data) did not exist previously. Imagine for example that the end user doesn't need arbitrary terrestrial weather data covering Florida but rather needs an example of a typical Florida spring storm. The ESG was developed to address this problem. Because the functionality of the ESG mimics that typically performed by a human expert, it was natural to turn to the field of computer intelligence in our search for a solution. In particular, we looked at the broad topic of data mining as outlined in (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Another prime requirement of the ESG system design was to allow the ESG user to query the archives in human linguistic terms. Natural language is not easily translated into the absolute terms of 0 and 1, which make up the digital world. The mapping between human language and computer systems is typical of Fuzzy Logic systems. Fuzzy logic is a superset of conventional (Boolean) logic that has been extended to handle the concept of partial truth, i.e. values between "completely true" and "completely false". It was introduced by Dr. Lotfi Zadeh (Zadeh, 1965) of UC/Berkeley in the 1960's as a means to model the uncertainty of natural language. Some of the major advantages of taking a fuzzy based approach are: It allows more realistic (natural) definition of sets; it provides for more graceful handling of boundaries/intersections between sets; it provides more human-like searching than a classical approach (e.g., allowing linguistic terms like “very high” or low). The ESG architecture relies heavily on a Java based fuzzy logic engine to perform searching and analysis for the user.

The purpose of the ESG is fundamentally to help a user distill the vast amount of available data down to a manageable amount of information that is both relevant and appropriate to his needs. The researcher can use ESG to find out if and when a particular type of event occurs in a region, how often it might occur and what the trend has been for a given time period. This could be used, for example, in modeling communication, trafficability or emergency services in response to environmental events. Beyond this, however, the ESG has applications in the area of data quality control, data classification and even forecasting. The increasing volume of data available in the future demands different techniques to handle it, and the ESG framework presented below is one proven method for a data manager to cope with this problem.

In the sections below we map out an overview of the ESG system. In section two we discuss the software architecture of the system, in section three we go into some detail about the fuzzy logic system, in section four we discuss some of the current data sources, and in section five we present a case study. Like any major software system, the ESG is the composition of many software components, but we attempt to present the broad picture design of ESG in the hopes it will serve as a pattern for those faced with a similar challenge when handling their own distributed environmental archive challenges.

## 2 ESG ARCHITECTURE

The ESG is designed using a distributed N-tier pattern (Larman, 2002), which may be roughly divided into user interface (UI), services, and data sources layers. The ESG UI is implemented as a web application which interacts with the services and data sources and generates dynamic content using Java Server Pages (JSP) technology. Part of the interaction with a user, such as verification of web forms and animated plots of environmental data, is done in the client’s browser using Java applets and JavaScript. ESG services are a set of Java Servlets (Sun Microsystems, Inc., n.d.a) and JAX-RPC (Sun Microsystems, Inc.., n.d.b) web services (W3C, n.d.) activated at different steps of the system workflow that perform data discovery, collection, mining and modeling, visualization and mapping, encoding and delivery to the end user. ESG data sources are a set of environmental databases and repositories of transient environmental data files interfaced to the system through web services which all conform to the simple ESG Data API standard.

In the core of the ESG services layer we have a data discovery and collection service (referred to as the Data API), common terminology service, data mining service (referred to as the Fuzzy Engine), several data encoding services (referred to as Encoders), map service, and a visualization service. A user’s typical workflow through the ESG system (Figure 1) includes the dynamic discovery of data sources based on their metadata provided by the Data API, presentation of the spatial properties of the data source by the map service, composition of an environmental scenario using the scenario editor from the User Interface, a data order from the Data API, data mining by the Fuzzy Engine, presentation of the resulting scenarios through the visualization service, and finally encoding and delivery of the properly formatted data to the end user by the Encoder service. The common terminology service provides common names, units of measurement, short descriptions, and references to all the environmental parameters searchable by ESG.

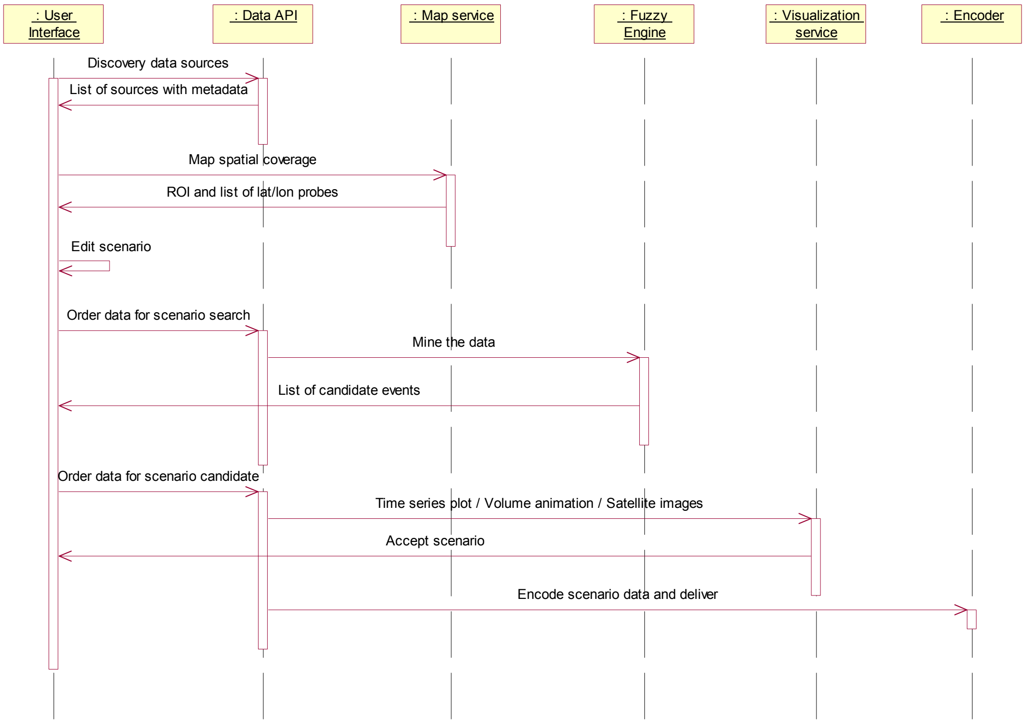

: User Interface
: Data API
: Map service
: Fuzzy Engine
: Visualization service
: Encoder
Discovery data sources
List of sources with metadata
Map spatial coverage
ROI and list of lat/lon probes
Edit scenario
Order data for scenario search
Mine the data
List of candidate events
Order data for scenario candidate
Time series plot / Volume animation / Satellite images
Accept scenario
Encode scenario data and deliver

**Figure 1.** Typical interactive ESG system workflow

One of the important design considerations in a distributed system is how data will propagate throughout the architecture. In the environmental domain several prominent data models exist such as netCDF (Unidata, n.d.a), HDF (NCSA, n.d.), VisAD (Hibbard, 1998), etc. However these data models are optimized for data exchange, whereas for ESG we needed a computationally optimized data model. Another problem affecting sophisticated scientific formats is versioning; e.g., the system designed for HDF version 4 will not process HDF version 5 data. The need for a small compact data model capable of easily interfacing with various service components and simple enough to be immune to version changes led us to design the *EsgPack* data model as a minimal object capable of transporting environmental data types but not adding any overhead which could reduce system performance. The *EsgPack* may be considered a JavaBean (Sun Microsystems, Inc., n.d.c) with only "basic" type properties such as string, integer, float array, which makes it possible to serialize it using standard JAX-RPC factories and to easily port between .NET services implemented in different languages such as Java, C#, or Perl.

To further simplify access to data throughout the system we chose to implement the Data API as a web-service (Figure 2). Web services are programmatic services offered via the Web. In a typical Web services scenario, an application sends a request to a service at a given URL using the SOAP protocol over HTTP. The service receives the request, processes it, and returns a response. The language neutral nature of the exchange means that the requesting program can be of any form and still use the service. In our application to get data from the web-service implementing the ESG Data API, one simply has to call the *getData()* method via JAX-RPC with several parameters, such as data source ID, parameter name, grid point coordinates, time interval, and output format. As a result of such a call (if successful), the web service returns a URL to the exported data file, which may be downloaded and processed by the calling client. The default output format is a serialized collection of *EsgPack* Java objects. The web service can also be used to export data in netCDF format. The metadata describing environmental parameters available through each data source, such as display names, height levels, units of measurement, scaling factors, etc., is provided by the same web service using the *getMetadata()* function, which receives a data source ID as the only input parameter and returns an XML *Element* object filled with metadata. Note that, this way a given environmental database may appear in ESG as several data sources under different IDs with different units of measurement for the same parameter (e.g., K and F degrees for temperature) and a different set of "virtual" parameters which are not stored but calculated on the fly at the time of the *getData()* function call from the parameters stored in the database (e.g., wind direction calculated from the U- and V-wind speed components). For transient data sources, such as an HDF or netCDF file, the metadata is dynamically discovered from the data file or application server at the time of the *getMetadata()* function call.

To run a fuzzy search on a data source with known ID, we simply have to call the web service *getData()* function, which will extract a part of the time series for the given time interval and combination of parameter/level/location and serialize it into the *EsgPack* object, then download the exported data from the URL returned by the web service call and pass it to the ESG fuzzy search engine.

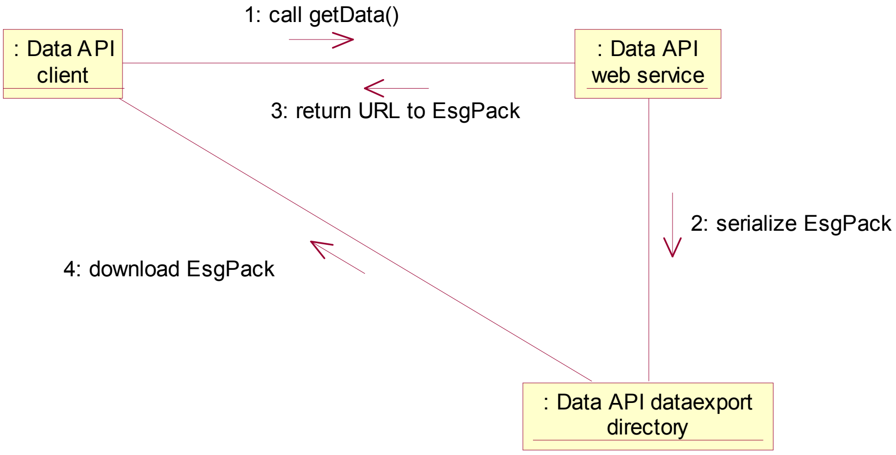
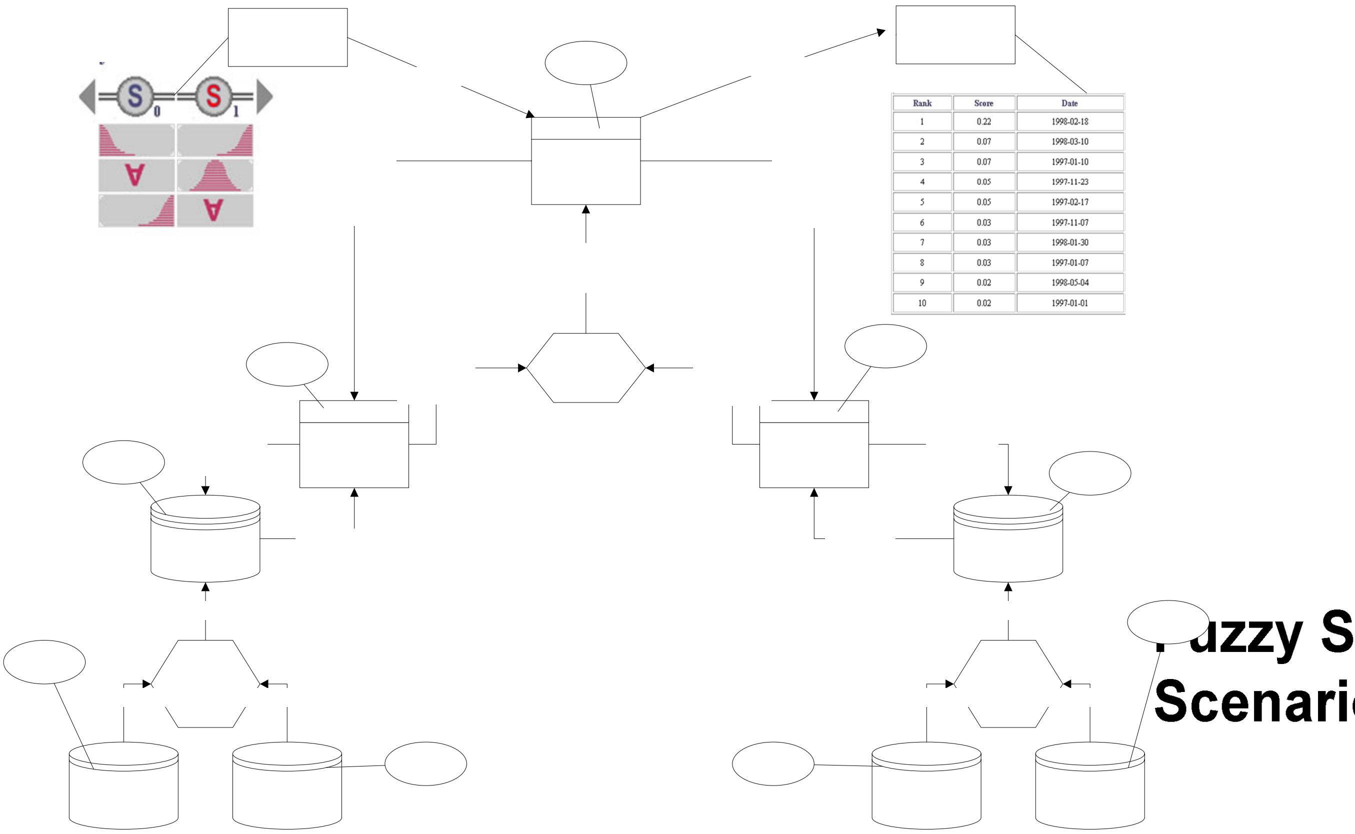


**Figure 2.** ESG Data API web-service *getData()* activity

The ESG system deployment diagram is a hierarchy of several layers of parallel computer clusters, with a cluster of application servers running ESG Services at the top connected to a set of database clusters serving as separate ESG data sources (Figure 3). For example (see the Data Sources section below), the National Centers for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR) Reanalysis Project (Kalnay, 1996) data source with global coverage of meteorological data for 1949-2000 is in fact a cluster of 5 parallel database servers (we call them local database servers, LDS) each loaded with 10 years of data. The National Geophysical Data Center (NGDC) Space Physics Interactive Data Resource (SPIDR, n.d.) data store with space weather data for 1933-2002 is a cluster of 6 parallel database servers each storing a subject-specific database, e.g., geomagnetic variations or Geostationary Operational Environmental Satellite (GOES) measurements. ESG Data API services to access each of these two data sources reside in a separate web service container (we call it the global database server, GDS) in the cluster of ESG application servers. To data mine an environmental scenario that includes a combination of meteorological and space weather factors, we can then run parallel searches on the 2 application servers with Fuzzy Engines requesting data separately from NCEP/NCAR and SPIDR GDS's, and at the end merge the fuzzy scores of the candidate events on the application server dedicated to the User Interface. This architecture allows us to use the Fuzzy Engine to execute very large queries in an interactive fashion as opposed to long asynchronous runs.  In our feedback sessions with ESG system users we find they get far more out of their investigation of the data when it's done interactively allowing them to easily tune their queries.



| Rank | Score | Date |
|---|---|---|
| 1 | 0.22 | 1998-02-18 |
| 2 | 0.07 | 1998-03-10 |
| 3 | 0.07 | 1997-01-10 |
| 4 | 0.05 | 1997-11-23 |
| 5 | 0.05 | 1997-02-17 |
| 6 | 0.03 | 1997-11-07 |
| 7 | 0.03 | 1998-01-30 |
| 8 | 0.03 | 1997-01-07 |
| 9 | 0.02 | 1998-05-04 |
| 10 | 0.02 | 1997-01-01 |

**Figure 3.** ESG deployment is a hierarchy of several layers of parallel computer clusters. The example shown is for the case of two distributed database systems.

## 3 THE FUZZY LOGIC SYSTEM

A classical set *A* in a space of objects *X* (called the universe) can be defined by its indicator function $I_A(\cdot): X \to \{0,1\}$, which is equal to 1 for all elements *x* from the set A and to 0 otherwise. Here we plot an indicator function $I_{[5,8]}R \to \{0,1\}$ of the segment [5,8] as a subset of all real numbers $R = \left]-\infty,+\infty\right[$.

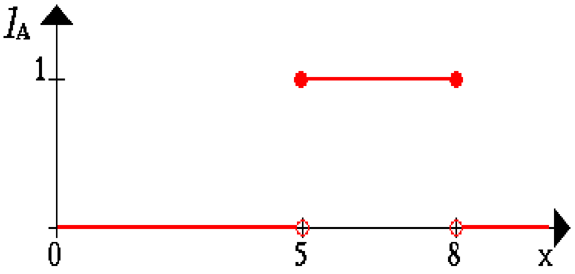


**Figure 4.** Indicator function for the set 5-8.

CPU

Data Request
probes
parameteres
time interval

CPU 4

A **fuzzy set** expresses degree to which an element belongs to a set. Hence the indicator function of the fuzzy set is allowed to have values between 0 and 1, which denotes degree of membership of an element in a given set.

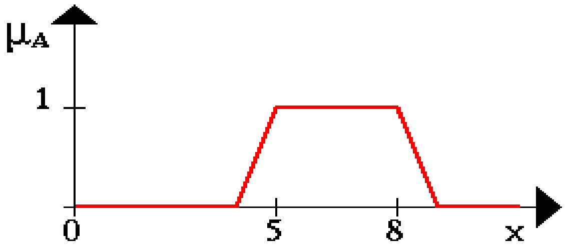


**Figure 5.** Fuzzy indicator function for the set 5-8.

A fuzzy set *A* in *X* is defined by its **membership function** (or MF for a short) $\mu_A(\cdot): X \to [0,1]$, which maps each element of *X* to its membership grade. Compare graphs of a MF for the fuzzy interval between 5 and 8 and the indicator function for the classical segment [5,8] above (Figures 4 & 5) .

Classical set theory / mathematical logic dualism (set union = logical OR, set intersection = logical AND, set complement = logical NOT) may be generalized for the fuzzy logic/set case in any different ways. One of the simplest generalizations for two fuzzy sets A and B is to use minimum of MFs $\min(\mu_A, \mu_B)$ for the intersection of fuzzy sets (fuzzy logic AND),

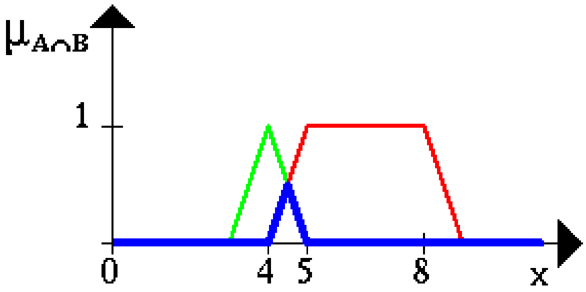


**Figure 6.** Fuzzy AND.

maximum of MFs $\max(\mu_A, \mu_b)$ for fuzzy set union (fuzzy logic OR),

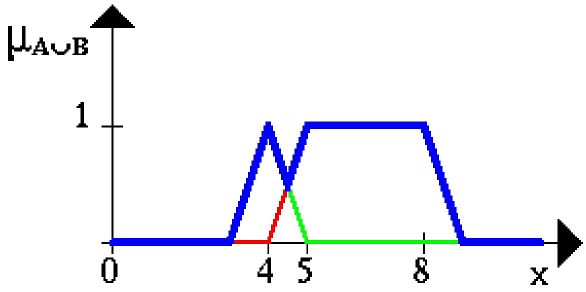


**Figure 7.** Fuzzy OR.

and $\mu_{\bar{A}} = 1 - \mu_A$ for fuzzy set complement (fuzzy logic NOT).

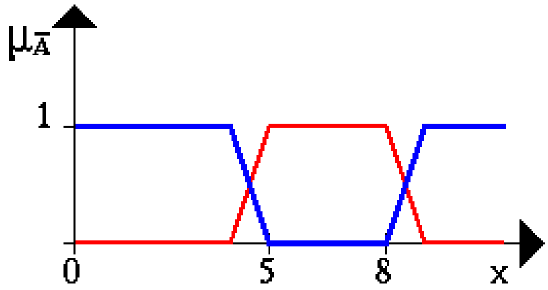


**Figure 8.** Fuzzy NOT.

People often use qualitative notions to describe such variables as temperature, pressure, pulse rate. In reality, it is difficult to put a single threshold between what is called "warm" and "hot". Fuzzy set theory serves as a translator from vague linguistic terms into strict mathematical objects.

Intelligent environmental scenario searching across the distributed resources is performed within the ESG's fuzzy search engine. The scenario editor from the ESG UI layer is used to formulate a set of conditions to be satisfied by the candidate events. The search conditions may be specified in a number of ways depending on the user's familiarity with the region/data of interest. An expert user can specify exact thresholds and/or limitations that must be maintained on certain parameters. Conditions can also be specified via abstract natural language definitions for each parameter. For instance, temperature limitations can be specified as “hot”, “cold”, or “typical”. The query can also be specified in terms of predefined rules which collect conditions into a named set. Thus, a user can specify the following fuzzy search request:

(VERY LARGE "precipitation rate") AND ("surface temperature" ABOUT 10°C) AND (LOW "vertical wind speed at pressure level 1000 mbar")

The result of such a request reported by the fuzzy search engine is always a list of the "most likely" dates for the event ranked by the sorted values of the aggregated multidimensional fuzzy membership function (MF).

Fuzzy search patterns in the multidimensional ESG databases are specified as logical AND aggregations of one-dimensional MFs formed for each variable (dimension) using the generalized bell function (Jang, Sun & Mizutani, 1997)

$$\mu_{gbell}(\tilde{x}; a, b, c) = \frac{1}{1 + \left| \frac{\tilde{x} - c}{a} \right|^{2b}}$$

Here, $\tilde{x}$ stands for normalized for range [0,1] scalar data variable, *c* stands for center of the symmetrical "bell", *a* for its half-width, and *b/2a* controls its slope. We use here simple range normalization for the variable $x_{\min} \le x \le x_{\max}$ : $\tilde{x} = (x - x_{\min})/(x_{\max} - x_{\min})$.

Five one-dimensional ESG fuzzy membership functions for a linguistic term set {very small, small, average, large, very large} are plotted below (Figure 9). Center, slope, and half-width of the bell functions for these linguistic terms are

**Table 1.** Linguistic term mappings.

| Linguistic term | Center | Slope | Half-width |
|---|---|---|---|
| very small | 0 | 5 | 0.2 |
| Small | 0.25 | 5 | 0.2 |
| Average | 0.5 | 5 | 0.2 |
| Large | 0.75 | 5 | 0.2 |
| very large | 1 | 5 | 0.2 |

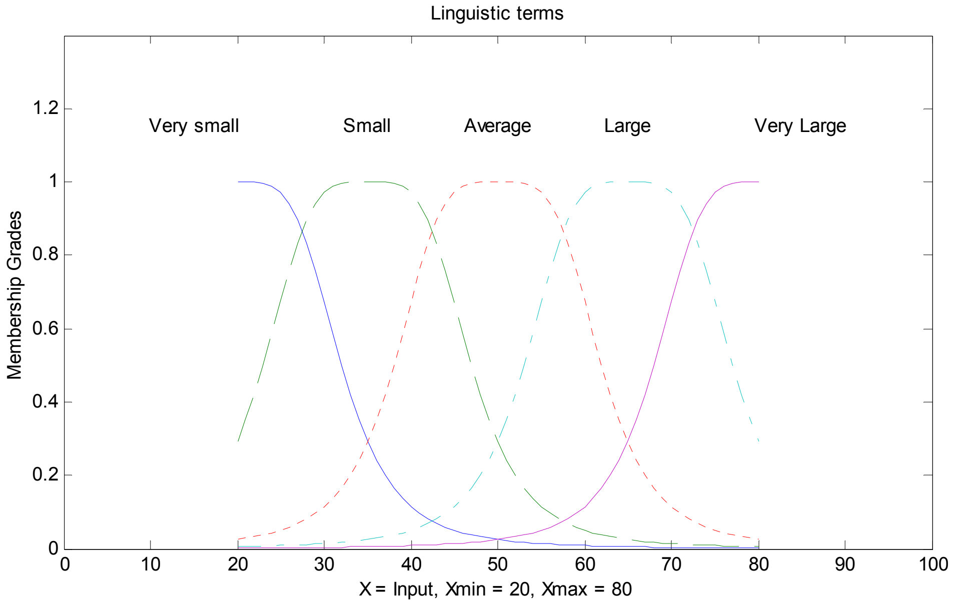


**Figure 9.** Linguistic to numeric translation of fuzzy terms.

On the next graph (Figure 10) we present examples of four one-dimensional MFs from the ESG numerical fuzzy term set {less than, about, between, greater than}. Let the variable $x$ vary in the range $x_{\min} \leq x \leq x_{\max}$ , then the center, slope, and half-width of the bell functions for numerical terms are

**Table 2.** Term mapping functions.

| Numerical term | Center | Slope | Half-width |
|---|---|---|---|
| less than *X*, $X < x_{\min}$ | $(X - x_{\min})/(x_{\max} - x_{\min})$ | 10 | 0.1 |
| less than *X*, $x_{\min} \leq X$ | 0 | 10 | $(X - x_{\min})/(x_{\max} - x_{\min})$ |
| about *X* | $(X - x_{\min})/(x_{\max} - x_{\min})$ | 10 | 0.1 |
| between *X* and *Y* | $((X + Y)/2 - x_{\min})/(x_{\max} - x_{\min})$ | 10 | $(Y - X)/2(x_{\max} - x_{\min})$ |
| greater than *X*, $X \leq x_{\max}$ | 1 | 10 | $1-(X - x_{\min})/(x_{\max} - x_{\min})$ |
| greater than *X*, $x_{\max} \leq X$ | $(X - x_{\min})/(x_{\max} - x_{\min})$ | 10 | 0.1 |

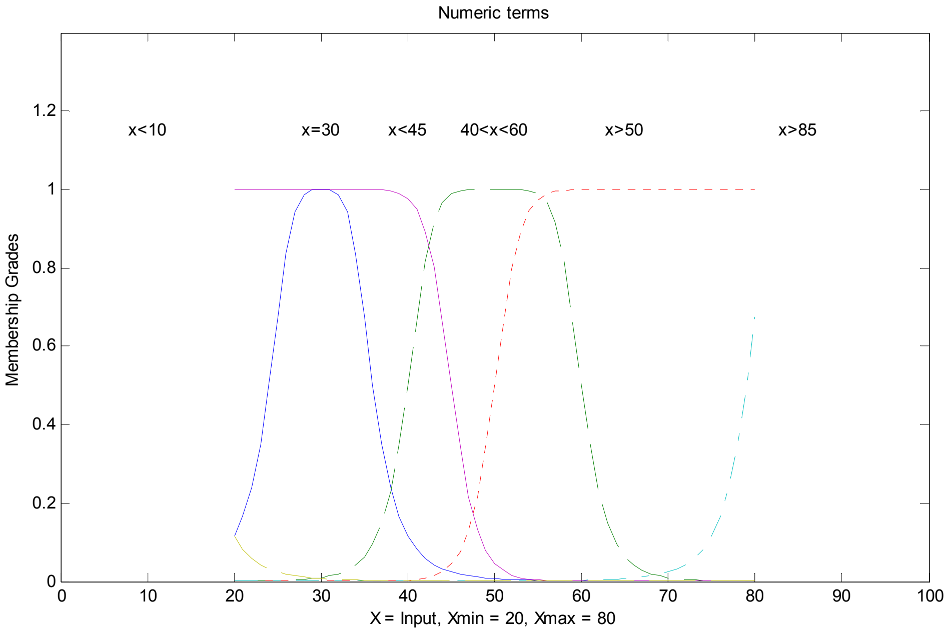


**Figure 10.** One dimensional fuzzy membership functions

Performing a Fuzzy 'AND' aggregation of two one-dimensional MFs $\mu_A(x), \mu_B(y)$ is conducted by Yager's (Yager 1980) T-norm operator (we use $q$=5):

$$T_Y(\mu_A(x), \mu_B(y), q) = 1 - \min\left\{1, \left[(1-\mu_A(x))^q + (1-\mu_B(y))^q\right]^{1/q}\right\} \quad q \geq 1$$

The formula of the Yager's T-norm operator for fuzzy AND aggregation of any $N$>1 one-dimensional MFs $\mu_n(x),\ n = 1 \ldots N$ is (Jang et al., 1997)

$$T_Y(\mu_n(x), q) = 1 - \min\left\{1, \left[\sum_{n=1}^{N}(1-\mu_n(x))^q\right]^{1/q}\right\}, \quad q \geq 1$$

The resulting surface of values for the multi-dimensional MF is more smooth than using a simple minimum of the aggregating MFs, which is the limit case of Yager's T-norm for q=1. To illustrate this we plot in Figure 11 an example of two trapezoidal membership functions aggregated by a simple MIN function, two bell membership functions aggregated by MIN and two bell membership functions aggregated by Yager's T-norm operator with q = 5. Here we are presenting the result for the fuzzy version of (-5<X<5) AND (-5<Y<5).

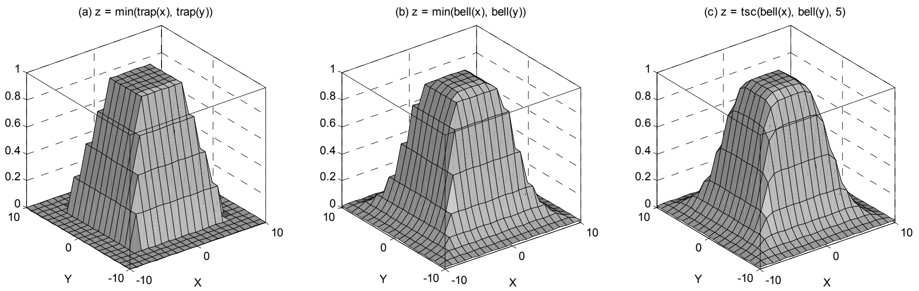


**Figure 11.** Examples of multi-dimension membership aggregation

In the ESG application we are searching for events in the environment where the input variables and the one-dimensional MFs $\mu_n(x(t))$ depend on time, as well as the fuzzy AND aggregation of the desired conditions $T_Y(\mu_n(x(t)),q)$ . We consider the values of the resulting time series $T_Y(\mu_n(x(t)),q)$ as the "likeliness" of the environmental event to occur at the time moment *t*, and search for the highest values of the aggregated MF and consider these to be the most likely candidates of the environmental events.

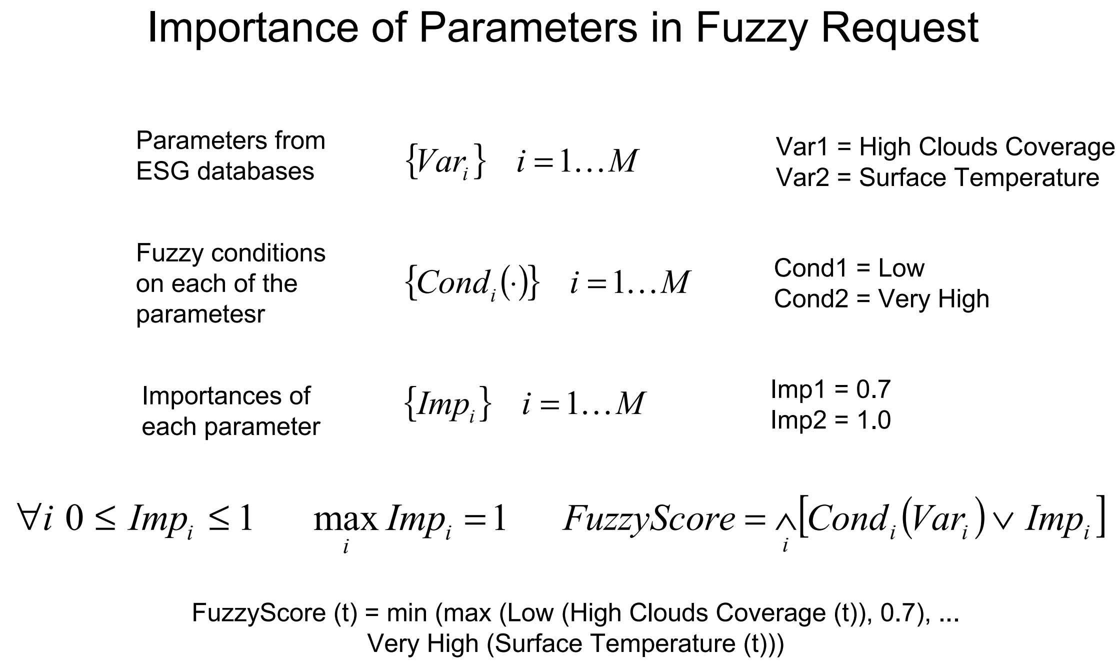
We use a simple climatology analysis to obtain normalization limits in Table 2. The limits are set to the minimum and maximum parameter values observed within the continuous or seasonal intervals given by the time constraints of the fuzzy search.

To be able to search for events like "the hottest day" or "the hottest week" we introduce the concept of an *event duration* $k\Delta t$ which may be any multiple $k = 1, 2, ...$ of the time step of the input variables $\Delta t$ . For example, the time step of the parameters from the NCEP/NCAR reanalysis database is 6 hours, so the minimum event duration is also 6 hours, but the event duration may be also 1 day, 1 week, etc. We do a moving average of the input parameters with the time window of the event duration before calculation of the one-dimensional MFs and the fuzzy AND aggregation:

$$\bar{x}(t_i) = \frac{1}{k}\sum_{j=i}^{i+k-1} x(t_j), \quad t_i = t_0 + i\Delta t .$$

To search for "the hottest day", first we have to smooth the air temperature with the time window of 1 day ( $k = 4$ ) and then take the time when the maximum of the smoothed temperature has occurred. To have the result of the fuzzy search in the form of a ranked list of the *K*-most likely dates (times) of the events, we sort the aggregated MF $T_Y(\mu_n(\bar{x}(t_i)),q)$ and select the times of the *K* maximum values of the aggregated MF $\{t_{i_1}, t_{i_2}, ..., t_{i_K}\}$ separated in time by the intervals longer than the event duration: $t_{i_{l+1}} - t_{i_l} > k\Delta t$ .

The fuzzy search request may contain conditions which never or very rarely take place at the same time at the specified location, although they can be observed there separately at different time moments. For example, very high precipitation rate and very high air pressure are unlikely to occur simultaneously. The fuzzy search for such a combination of conditions may return an empty set of candidate dates and times. We decrease the probability of the empty fuzzy search results by introducing the concept of *importance* of the input parameters. The importance $w_n$ is a constant weight of a given parameter in the range between 0 and 1. More important parameters are given higher weight, with the condition that the highest priority is then normalized to one. Then instead of the one-dimensional MFs $\mu_n(x(t_i))$ in the fuzzy AND aggregation we use “optimistic” values $\max[\mu_n(x(t_i)), w_n]$. For parameters with importance 1 we use the original MFs as before, and the parameters with the importance 0 are not used in the search at all.



Importance of Parameters in Fuzzy Request

| | | |
|---|---|---|
| Parameters from ESG databases | $\{Var_i\} \quad i = 1 \ldots M$ | Var1 = High Clouds Coverage<br>Var2 = Surface Temperature |
| Fuzzy conditions on each of the parametesr | $\{Cond_i(\cdot)\} \quad i = 1 \ldots M$ | Cond1 = Low<br>Cond2 = Very High |
| Importances of each parameter | $\{Imp_i\} \quad i = 1 \ldots M$ | Imp1 = 0.7<br>Imp2 = 1.0 |

$$\forall i \; 0 \le Imp_i \le 1 \qquad \max_i Imp_i = 1 \qquad FuzzyScore = \wedge_i \left[ Cond_i(Var_i) \vee Imp_i \right]$$

FuzzyScore (t) = min (max (Low (High Clouds Coverage (t)), 0.7), ...
Very High (Surface Temperature (t)))

An application of the fuzzy engine is presented in the case study, Section 5 below.

## 4 DATA SOURCES

The connections between the ESG system and a given user community are the data sources that support it. As discussed above it’s relatively easy to hook new data sources to the ESG through the web-services interface so this list shouldn’t be taken as limiting but rather as a snapshot of current functionality. The table below is meant to give the reader an idea of how ESG is being used now and the types of data being accessed.

**Table 3.** ESG data sources.

| Data Source | Type | Sample Parameters | Temporal Coverage | Spatial Coverage | Size | URL |
|---|---|---|---|---|---|---|
| NCEP/NCAR | Meteo. | Temperature, Windspeed, Cloud Cover | 1949 – 2000 | Global @2.5 Deg. | 250 Gb | http://dss.ucar.edu/pub/reanalysis/ |
| ACMES | Meteo. | Temperature, Sea State, Snow Cover | Various | Regional @ 10 and 40 km | 20 Gb / region | https://msea.afccc.af.mil/html/projects.html#acmes |
| SWR | Space | Electric Potential, Auroral Precipitation | 1997 - 1998 | Global Mixed resolution | 3 Tb | http://swr1.ngdc.noaa.gov/swr |
| SPIDR | Space | Kp index, Sunspot Number | 1933 – 2002 | Global | 30 Gb | http://spidr.ngdc.noaa.gov |

The first thing to notice from above is the relatively large size of the archives. Using the distributed database concept allows us to perform interactive mining on these substantial data sources. The second thing to notice is the long temporal ranges represented. The ESG is most useful when the size of the archive prohibits or makes searching by hand impractical. To describe each resource briefly, both the Advanced Climate Modeling Environmental Simulations (ACMES) and NCEP/NCAR data archives are derived from numerical weather prediction model runs. They represent gridded output on a regular time step and fixed grid. The models use data ingest procedures to assimilate observational data into model results to produce a consistent picture of the environment. They are excellent data sources in that the parameters and observation points are fixed making analysis relatively easy. The problem is they tend to be coarse with 1.0 and 2.5 degree grid steps respectively. When higher resolution data is required the ESG project typically generates runs for specific regions and then registers those data sources for use by the system.

The Space Weather Reanalysis (SWR) archive is a complete space weather representation using physically consistent data-driven space weather models. The product is a consistent, integrated historical record of the near Earth space environment obtained by coupling observational data from space environmental monitoring system archives with data-driven, physically based numerical models. The resulting database is an enhanced look at the space environment on consistent grids, time resolution, coordinate systems and containing key fields allowing a modeling and simulation customer to quickly and easily incorporate the impact of the near-Earth space climate in environmentally sensitive models. Similar to the numerical weather prediction models described above, the SWR data is a coarse representation of the domain but it can often be used to initialize production of data at higher resolutions.

The last data source, the Space Physics Interactive Data Resource (SPIDR), is unique in that it is an observational data source and not the output of models. It therefore makes no guarantee of data continuity or coverage, unlike the other resources. The SPIDR system currently handles the following: Defense Meteorological Satellite Program (DMSP) visible, infrared and microwave browse imagery, ionospheric parameters, geomagnetic 1.0 minute and hourly value data, geophysical and solar indices, GOES satellite x-ray, plasma, and magnetometer data, cosmic ray, solar radio telescope, satellite anomaly, and city lights data sets.

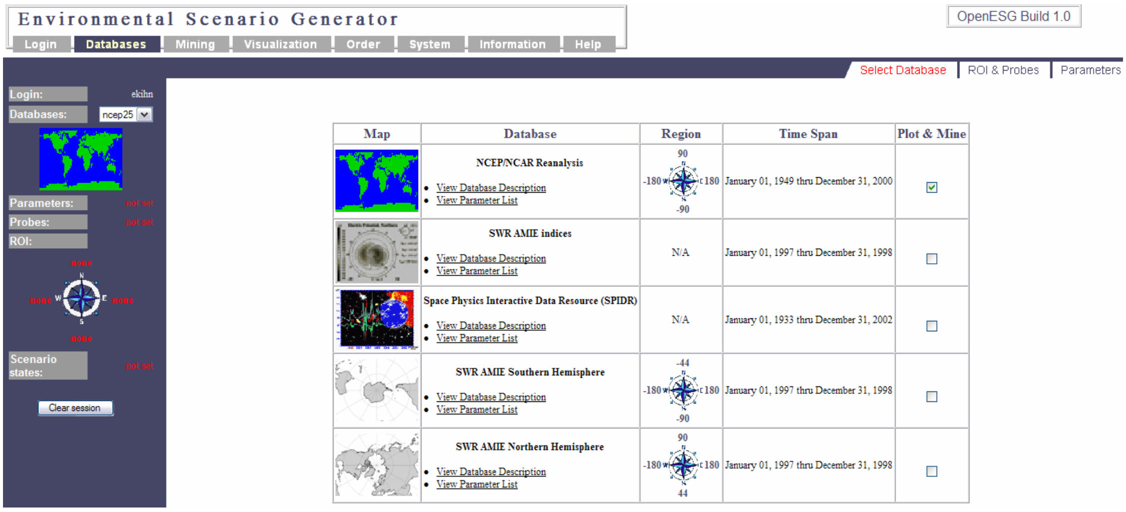
There are plans to add ocean and terrain data in the near future, making the ESG mining technology available across a wide representation of the Earth's environment.

## 5 CASE STUDY

This section presents an account of the use of the ESG in support of a virtual demo of the Grizzly, a prototype mine clearing, tracked vehicle. The Grizzly project asked the ESG team to find a very wet period followed by immediate drying conditions for a specified region of interest (ROI) in the National Training Center (NTC). The ROI specified was:

35 deg 44 min N to 34 deg 15 min N, 116 deg 59 min W to 115 deg 45 min W

This scenario was run to get environmental input for a trafficability model of the Grizzly. We began as in Figure 12 by selecting a data resource. In this case the choice of database was simple because we were looking for a terrestrial weather event and the NCEP/NCAR database contains parameters that are of direct interest to our study.

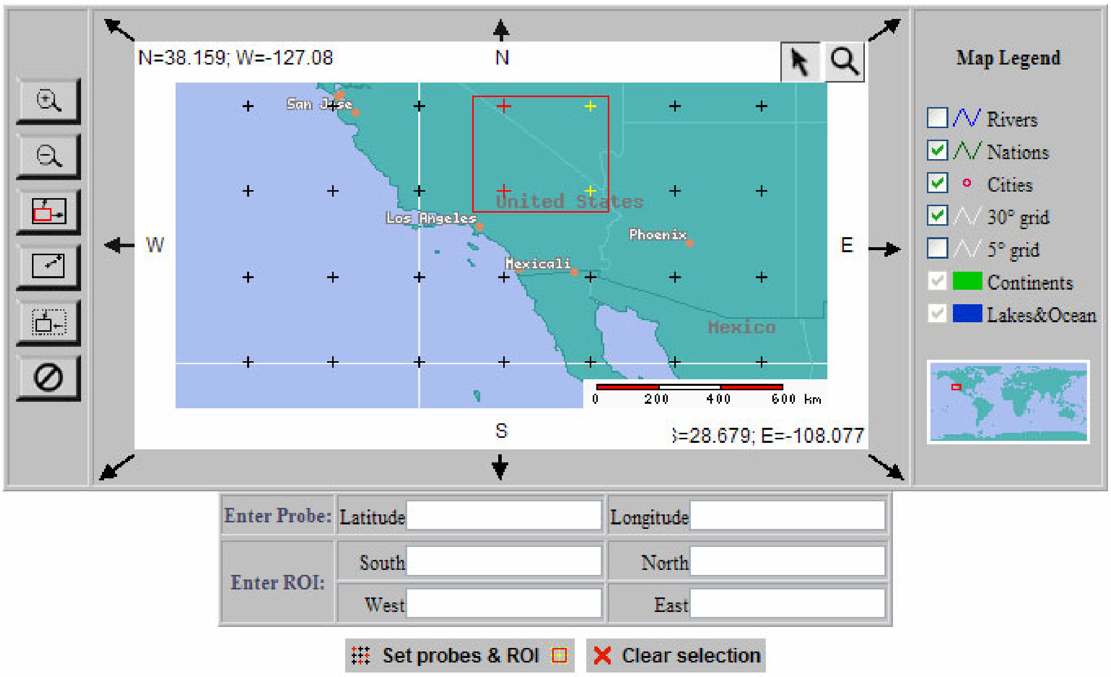


Login: ekihn
Databases: ncep25
Parameters: not set
Probes: not set
ROI:
none
none W E none
none
Scenario states: not set

| Map | Database | Region | Time Span | Plot & Mine |
|---|---|---|---|---|
| | NCEP/NCAR Reanalysis<br>• View Database Description<br>• View Parameter List | 90<br>-180 W E 180<br>-90 | January 01, 1949 thru December 31, 2000 | ☑ |
| | SWR AMIE indices<br>• View Database Description<br>• View Parameter List | N/A | January 01, 1997 thru December 31, 1998 | ☐ |
| | Space Physics Interactive Data Resource (SPIDR)<br>• View Database Description<br>• View Parameter List | N/A | January 01, 1933 thru December 31, 2002 | ☐ |
| | SWR AMIE Southern Hemisphere<br>• View Database Description<br>• View Parameter List | -44<br>-180 W E 180<br>-90 | January 01, 1997 thru December 31, 1998 | ☐ |
| | SWR AMIE Northern Hemisphere<br>• View Database Description<br>• View Parameter List | 90<br>-180 W E 180<br>44 | January 01, 1997 thru December 31, 1998 | ☐ |

**Figure 12.** The data source selection menu

Next as shown in Figure 13 we navigated to the region of interest. The ESG Map Service drew the map and displayed selected geographical information such as cities and nations as an overlay. The crosses indicate the precise points where the database has data, in this case on a 2.5 degree grid mesh. After selecting two probes (shown in red in Figure 13) we selected specific parameters from the database we wished to work with. In this case the metadata describing the parameters was retrieved through the Data API as described above and the GUI was populated for us to make a selection. In our case we chose precipitation rate and wind speed, which is a virtual parameter (i.e., not stored but computed on the fly) for the first scenario segment (very wet period). For the second segment (the drying phase) we chose temperature, precipitation rate, and wind speed.

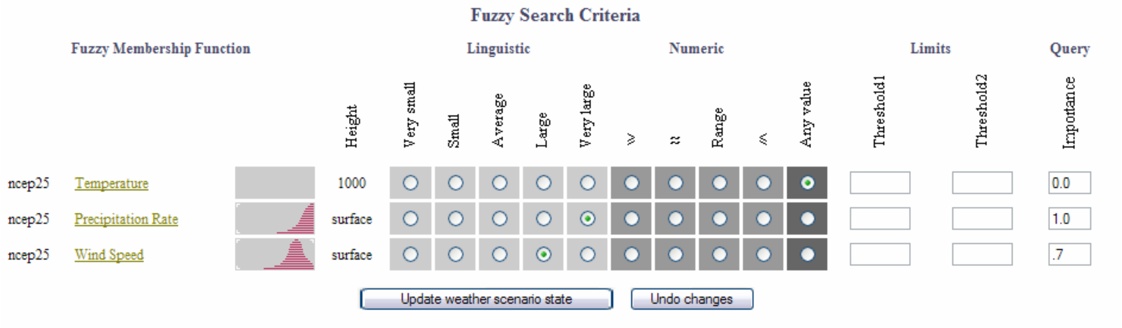


☐ Rivers
☑ Nations
☑ Cities
☑ 30° grid
☐ 5° grid
☑ Continents
☑ Lakes&Ocean

| | | | | |
|---|---|---|---|---|
| Enter Probe: | Latitude | | Longitude | |
| Enter ROI: | South | | North | |
| | West | | East | |

**Figure 13.** The spatial criteria screen

These parameters can be used to describe a very high rain event followed by a drying period of just the type we were looking for. Next the user must specify criteria on these parameters for each scenario segment used to model the type of event they are searching for. As shown in Figure 14 the criteria can be specified in human linguistic terms such as “Very Large” or “Small” or in numeric terms such as “greater than” or in a defined range. Since we didn’t know anything about the climatology of the parameters for the two probes of interest, we specified “Very Large” for precipitation rate and “Large” for wind speed. Notice how this helps us if we need to repeat the search in another region where, for example, “Very Large” wind speed might be quite different. The ESG computes, on the fly, the climatology for each region and time and handles the interpretation automatically. Notice also that we weighted our importance of parameters towards Precipitation Rate for this first segment. That emphasized the need for high rain-fall over wind conditions.

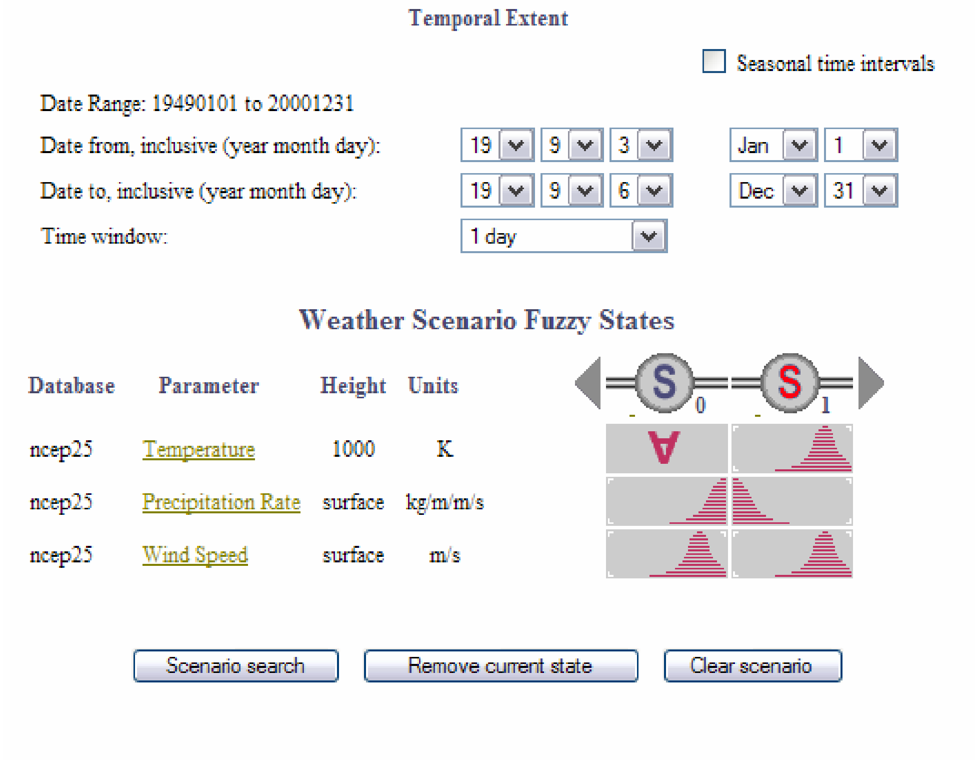


Fuzzy Search Criteria

| | Fuzzy Membership Function | | | Linguistic | | | | | Numeric | | | | | Limits | | Query |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Height | Very small | Small | Average | Large | Very large | ≥ | ≈ | Range | ≤ | Any value | Threshold1 | Threshold2 | Importance |
| ncep25 | Temperature | | 1000 | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ● | | | 0.0 |
| ncep25 | Precipitation Rate | | surface | ○ | ○ | ○ | ○ | ● | ○ | ○ | ○ | ○ | ○ | | | 1.0 |
| ncep25 | Wind Speed | | surface | ○ | ○ | ○ | ● | ○ | ○ | ○ | ○ | ○ | ○ | | | .7 |

**Figure 14.** Fuzzy search criteria specification

We repeated the specification for our second segment, this time emphasizing temperature and wind speed over precipitation rate. Next as shown in Figure 15 we specified the time range to search for the event described above. In our case we were required by outside conditions to use data from 1993 –1996, so we set a shorter search interval than is available through the data source.



Temporal Extent

☐ Seasonal time intervals

Date Range: 19490101 to 20001231

Date from, inclusive (year month day): 19 9 3 Jan 1

Date to, inclusive (year month day): 19 9 6 Dec 31

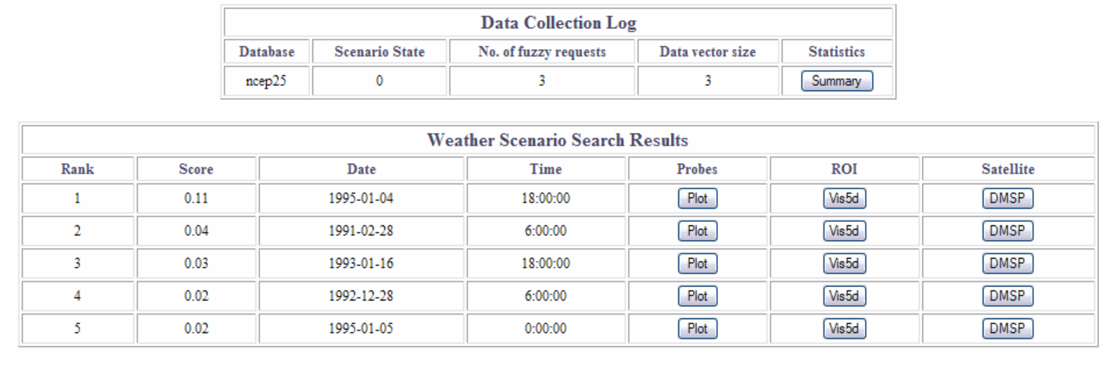
Time window: 1 day

Weather Scenario Fuzzy States

| Database | Parameter | Height | Units |
|---|---|---|---|
| ncep25 | Temperature | 1000 | K |
| ncep25 | Precipitation Rate | surface | kg/m/m/s |
| ncep25 | Wind Speed | surface | m/s |

**Figure 15.** Interface for search range and segment composition.

A few seconds after pressing the “Scenario search” button (it typically takes seconds) a result appeared as in Figure 16. This result presents a list of likely events identified by their starting date and score. The score is an indicator from 0 – 1 of how well this event matches the request. The more parameters and points involved, typically the lower the score, but the best matches are presented for user review.



Data Collection Log

| Database | Scenario State | No. of fuzzy requests | Data vector size | Statistics |
|---|---|---|---|---|
| ncep25 | 0 | 3 | 3 | Summary |

Weather Scenario Search Results

| Rank | Score | Date | Time | Probes | ROI | Satellite |
|---|---|---|---|---|---|---|
| 1 | 0.11 | 1995-01-04 | 18:00:00 | Plot | Vis5d | DMSP |
| 2 | 0.04 | 1991-02-28 | 6:00:00 | Plot | Vis5d | DMSP |
| 3 | 0.03 | 1993-01-16 | 18:00:00 | Plot | Vis5d | DMSP |
| 4 | 0.02 | 1992-12-28 | 6:00:00 | Plot | Vis5d | DMSP |
| 5 | 0.02 | 1995-01-05 | 0:00:00 | Plot | Vis5d | DMSP |

**Figure 16.** A typical results list

In order to investigate the results, the ESG Visualization tools allowed us to plot the data as shown in Figure 17, create spatial maps of the data (not shown), and compare the data with satellite imagery of the place and time as shown in Figure 18.  In this case notice that both the raw data values and the satellite imagery serve well to describe a scenario of very wet weather at the locations we chose. This leads to potentially interesting uses of the ESG, e.g., searching satellite imagery archives not by place and time but by conditions of the observation, such as “cloud free” or “stormy”.

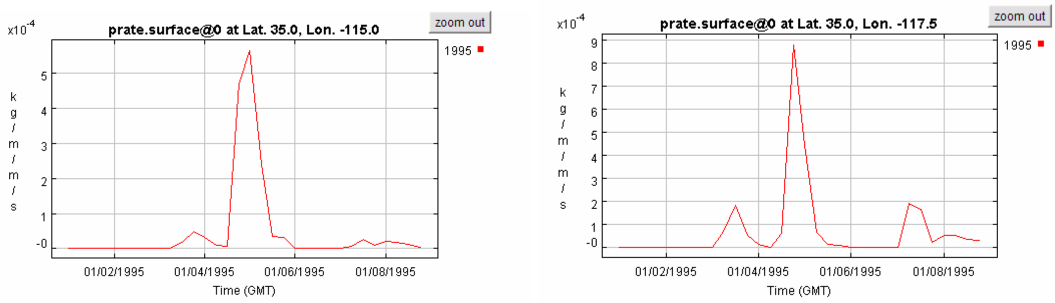


**Figure 17.** Plot of the data from the detected event.

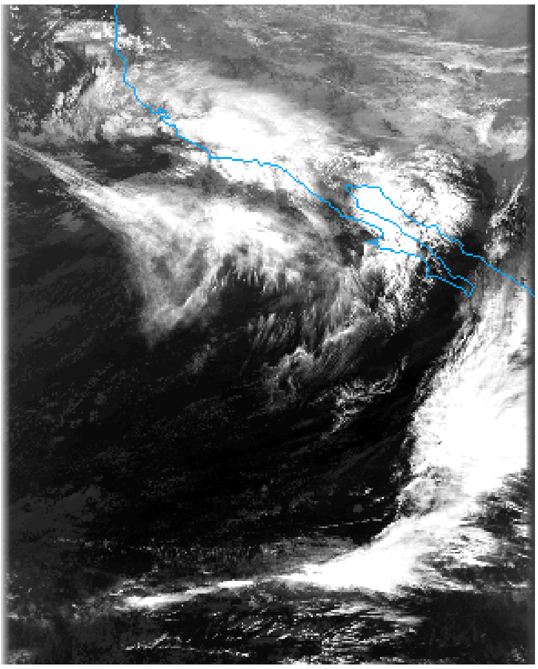


**Figure 18.** Satellite imagery showing observations of the event selected.

Finally once we reviewed the event and decided it met the criteria of the search, we downloaded and delivered the data for the entire region so that the Grizzly project could run with exactly the environmental scenario they needed. In the end it turned out to be a very powerful validation of the ESG technology because the event identified was a hurricane remnant passing over the region, which is a very rare event indeed for that region of interest, and it was automatically pulled from years of data and delivered to a happy user.

## 6 CONCLUSIONS

The applications of the ESG technology and ESG-like systems are broad. As more and more data archives become available through projects like CLASS (NOAA, n.d.), EOSDIS (NASA, n.d.), DODS (Unidata, n.d.b) and other network accessible data systems, the tools to extract information from them become more valuable. As Nature declared in a 1999 article (Reichhardt, 1999) "It's sink or swim as a tidal wave of data approaches**".** ESG can help users prepare by providing tools which sift through the vast quantities of data available online and point at the interesting bits. This means that even with the volume of data increasing so rapidly and the number of researchers remaining relatively level, we can hope to extract the most valuable information from the observations and model results and carry that back to the relevant scientific communities.

The application of fuzzy logic based data tools goes far beyond simple event selection. For example an ever present issue when dealing with these large data sets is quality control. There is simply too large a volume to reasonably screen by hand. Using techniques such as peer-matching and expert systems, we can extend the ESG to monitor data and alert data managers to changes and anomalies. This system is currently being used at NGDC for example in the quality control of geomagnetic station data where the system monitors that "like" stations see the same results during magnetic storms. As the computational power available expands we can extend the system into areas such as data classification, whereby we can identify modes of the environment and perhaps identify new unknown relations in specific regions.

Finally the emergence of a network infra-structure for data access is providing new opportunities for the scientific researcher. It is now fairly trivial to reach out across discipline boundaries and access data in an immediately useable format. This is true for example in the case of the terrestrial weather community being able to make use of the space data made available through the SWR. With these opportunities come challenges. As researchers expand into domains in which they may not be expert they will come to rely on intelligent tools to support them. Systems such as the ESG provide a pathway to capture "expert" domain knowledge and present it in an automated fashion to help just these users.

The interested reader can access the ESG system at http://esg.ngdc.noaa.gov

## 7 ACKNOWLEDGMENTS

The authors wish to thank the Defense Modeling and Simulation Office (DMSO) and the Air and Space Natural Environment (ASNE) Modeling & Simulation Executive Agent (MSEA) for their support of the work presented in this paper.

## 8 REFERENCES

DMSO (2003) Homepage of the Master Environmental Library.
Retrieved December 1, 2003 from the *DMSO* website: http://mel.dmso.mil

Fayyad, U.M., Piatetsky-Shapiro, G., & Smyth, P. (1996) From data mining to knowledge discovery: An overview. *Advances in Knowledge Discovery and Data Mining (*Chapter 1, 1-34). Menlo Park, CA: AAAI Press and the MIT Press.

Hibbard, W. (1998) VisAD: Connecting people to computations and people to people. *Computer Graphics 32*( 3), 10 -12.

Hilty, L.M., Page, B., Radermacher, F.J., & Riekert, W.-F. (Eds.)(1995) Environmental informatics as a new discipline of applied computer science. In Avouris, N.M. & Page, B. (Eds.) *Environmental Informatics - Methodology and Applications of Environmental Information Processing*. Norwell, MA: Kluwer Academic Publishers.

Jang, J.-S.R., Sun, C.-T., & Mizutani, E. (1997) *Neuro-Fuzzy and Soft Computing*. Upper Saddle River, NJ: Prentice Hall.

Kalnay, E.E.A. (1996) The NCEP/NCAR 40-year reanalysis project. *Bull Am. Meteorol. Soc. 77*, 437 - 471.

Larman, C. (2002) *Applying UML and Patterns, An Introduction to Object-Oriented Analysis and Design and the Unified Process.* Upper Saddle River, NJ: Prentice Hall PTR.

National Oceanic and Atmospheric Administration (NOAA)(n.d.) Homepage of the Comprehensive Large Array-data Stewardship System (CLASS). Retrieved December 1, 2003 from the *National Oceanic and Atmospheric Administration* website: http://www.saa.noaa.gov/cocoon/nsaa/products/welcome

National Aeronautics and Space Administration (n.d.) Homepage of the EOSDIS Core System (ECS) project.
Retrieved December 1, 2003 from the *National Aeronautics and Space Administration* website: http://edhs1.gsfc.nasa.gov/

National Center for Supercomputing Applications (n.d.) Homepage of the Hierarchical Data Format (HDF). Retrieved December 1, 2003 from the *National Center for Supercomputing Applications* website: http://hdf.ncsa.uiuc.edu/

NOAA Satellite & Information (n.d.) Homepage of the National Virtual Data System. Retrieved from the *NOAA Satellite & Information* website: http://nndc.noaa.gov/index.shtml

Reichhardt, T. (1999). It's sink or swim as a tidal wave of data approaches. *Nature 399,* 517-520.

Sun Microsystems, Inc. (n.d.a), Homepage of the Java 2 Enterprise Edition (J2EE).
Retrieved December 1, 2003 from the *Sun Microsystems, Inc.* website: http://java.sun.com/j2ee/

Sun Microsystems, Inc. (n.d.b) Homepage of the Java API for XML-Based RPC (JAX-RPC).
Retrieved December 1, 2003 from the *Sun Microsystems, Inc.* website: http://java.sun.com/xml/jaxrpc/

Sun Microsystems, Inc. (n.d.c), Homepage of the JavaBeans Spec.
Retrieved December 1, 2003 from the *Sun Microsystems, Inc* website.:
http://java.sun.com/products/javabeans/docs/spec.html

Unidata (n.d.a) Homepage of netCDF. Retrieved December 1, 2003 from the *Unidata* website:
http://www.unidata.ucar.edu/packages/netcdf/

Unidata (n.d.b) Homepage of the *DODS---Data Access Protocol*,
Retrieved December 1, 2003 from the *Unidata* website: *http://www.unidata.ucar.edu/packages/dods/*

World Wide Web Consortium (W3C) (n.d.) Web Services Activity.
Retrieved December 1, 2003 from the *World Wide Web Consortium* website: http://www.w3.org/2002/ws/

Yager, R. R. (1980) On a general class of fuzzy connectives. *Fuzzy Sets and Systems 4*, 235-242.

Zadeh, L. (1965) Fuzzy Sets. *Information and Control 8,* 338-353.

# 9 Appendix A

## Glossary of Terms

| | |
|---|---|
| ACMES | - Advanced Climate Modeling and Environmental Simulations |
| API | - Application Programming Interface |
| ASNE | - Air and Space Natural Environment |
| CLASS | - Comprehensive Large Array Stewardship System |
| CPU | - Central Processing Unit |
| DMSO | - Defense Modeling and Simulation Office |
| DMSP | - Defense Meteorological Satellite Program |
| DODS | - Distributed Oceanographic Data System |
| EOS | - Earth Observing System |
| EOSDIS | - EOS Data and Information System |
| ESG | - Environmental Scenario Generator |
| GDS | - Global Data Server |
| HDF | - Hierarchical Data Format |
| HTTP | - Hyper-Text Transfer Protocol |
| JAX-RPC | - Java API for XML-based RPC |
| JSP | - Java Server Pages |
| LDS | - Local Data Server |
| MEL | - Master Environmental Library |
| MF | - Membership Function |
| MSEA | - Modeling & Simulation Executive Agent |
| NCAR | - National Center for Atmospheric Research |
| NCSA | - National Center for Supercomputing Applications |
| NCEP | - National Centers for Environmental Prediction |
| netCDF | - Network Common Data Form |

| | |
|---|---|
| NASA | - National Aeronautics and Space Administration |
| NGDC | - National Geophysical Data Center |
| NOAA | - National Oceanic and Atmospheric Administration |
| NTC | - National Training Center |
| NVDS | - National Virtual Data System |
| ROI | - Region of Interest |
| RPC | - Remote Procedure Call |
| SOAP | - Simple Object Access Protocol |
| SPIDR | - Space Physics Interactive Data Resource |
| SWR | - Space Weather Reanalysis |
| UI | - User Interface |
| URL | - Uniform Resource Locator |
| W3C | - World Wide Web Consortium |