

# Meteorology and Space Weather Data Mining Portal

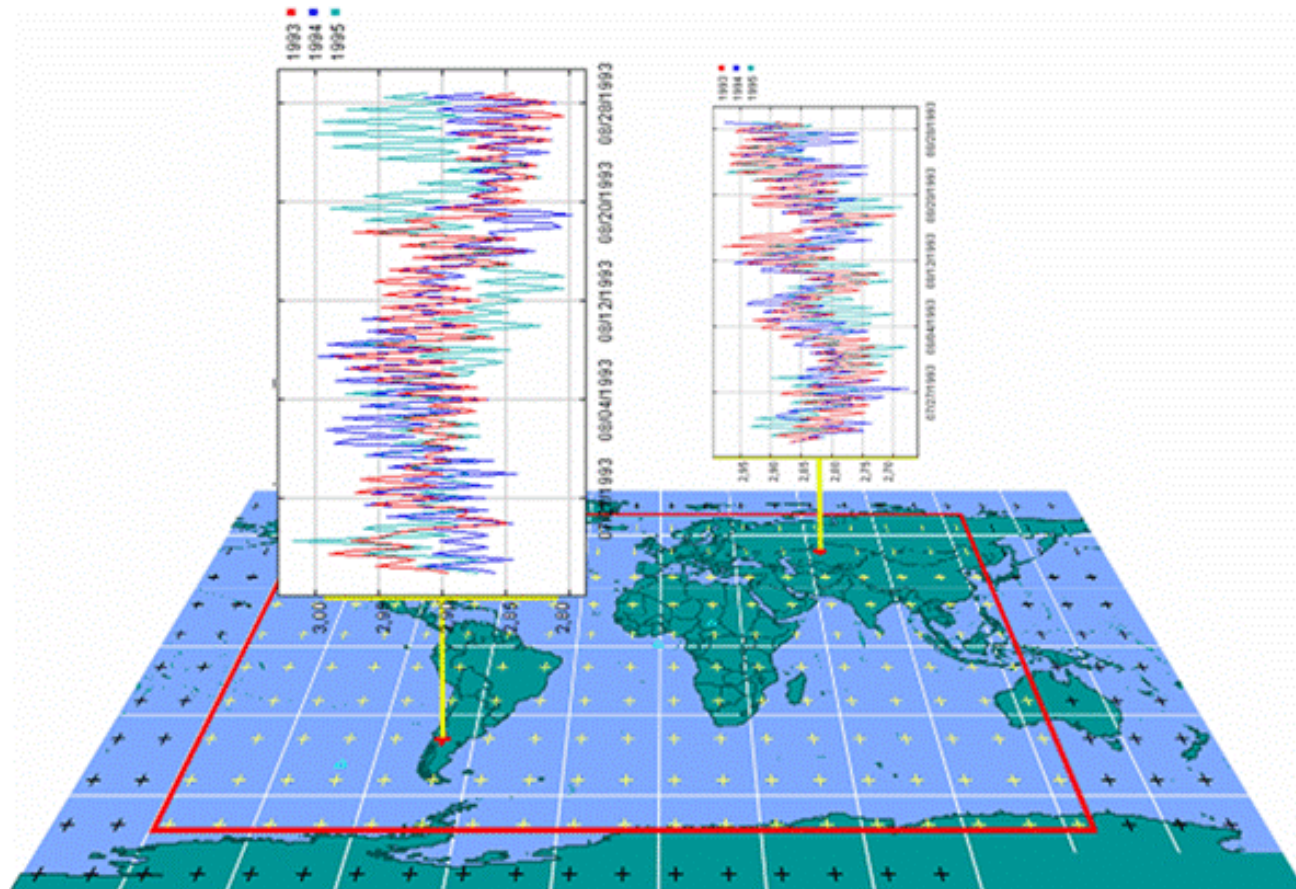
- **Dmitry MISHIN, Geophysical Center RAS**
- **Mikhail ZHIZHIN, Geophysical Center RAS**
- **Alexei POYDA, Moscow State University**

# Contents

1. Environmental data models
2. Metadata ordering and mining extensions
3. Supported data sources
4. Data mining extensions for OGSA-DAI
5. Environmental scenario defined by fuzzy logic
6. Data mining web portal workflow use case
7. Possible applications

# Environmental data models

Main environmental data structure is time series, i.e. an array of values of a parameter at different times on regular grid or specified locations (station data). Sequence of pairs, each having time and location is a trajectory.



# Metadata harvesting

IDEAS Integrated of Distributed Environmental Archives System

powered by

Login Databases Mining Visualization Documentation Help

Select Database | ROI and Probes | Parameters

Log in:  Admin

Databases:

Parameters:  set

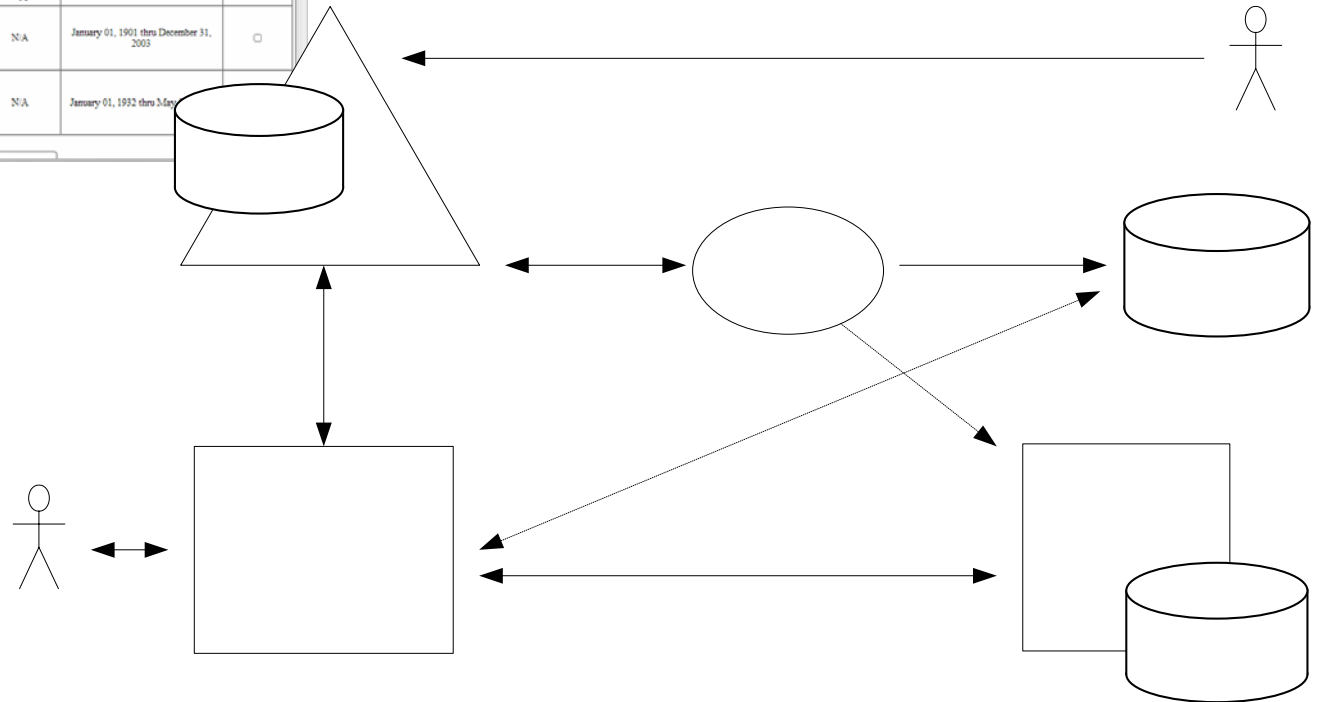
Probes:  set

ROI:

Scenario states:

Clear session

Map	Database	Region	Time Span	Plot & Mine
	NWS Weather Forecast (weather) • View Database Description • View Parameter List	90 -180 to 180 -90	January 01, 2005 thru December 31, 2006	<input type="checkbox"/>
	NCEP/NCAR Reanalysis (ncr25) • View Database Description • View Parameter List	90 -180 to 180 -90	January 01, 1949 thru December 31, 2006	<input checked="" type="checkbox"/>
	SPDR Geomagnetic Variations (hourly) [spidr_geom_hr] • View Database Description • View Parameter List	N/A	January 01, 1901 thru December 31, 2003	<input type="checkbox"/>
	SPDR Indices [spidr_indices] • View Database Description • View Parameter List	N/A	January 01, 1932 thru 3.1e...	<input type="checkbox"/>



# ES metadata ordering extensions

- Our metadata repository can handle different schemas in separate sections, f.e.:
  - FGDC
    - collection level, most suitable for digital maps, widely adopted by ES community
  - SPASE (NASA)
    - collection and inventory level, used by the Space Weather community
  - ECHO (NASA)
    - collection and inventory level, used by the Remote Sensing community
  - ESSE (NOAA and MSR)
    - collection and inventory level, used by the ESSE data mining project to describe virtual environmental data source in Grid
- Metadata ordering extensions are used to build a data request and fuzzy search for environmental scenario.

# Environmental data sources integration

## World Data Centers:

- SPIDR (Space Physics Interactive Data Archive)
  - From 1930 year
  - ~120 numerical parameters
  - ~0.5 TB

Space weather

## NOAA and ECMWF

- NCEP/NCAR Weather Reanalysis Project
  - From 1950 year
  - Weather parameters on regular grid, 2.5 deg step
  - ~1 TB
- ERA40 Weather Reanalysis Project
  - From 1957 year
  - Weather parameters on regular grid, 1 deg step
  - ~2 TB
- NWS Weather forecast
  - Weather parameters on regular grid, 1 deg step

Climatology models

## NOAA CLASS (Comprehensive Large Array-data Stewardship System)

- Satellite images
  - From 1992 year
  - Satellite images from ~100 spectral channels
  - ~1.2 PB, growing ~0.5 PB per year
- Time series data products

Remote sensing

OpenDAP servers network ...

# GRID data services:

<http://www.ogsadai.org.uk/>

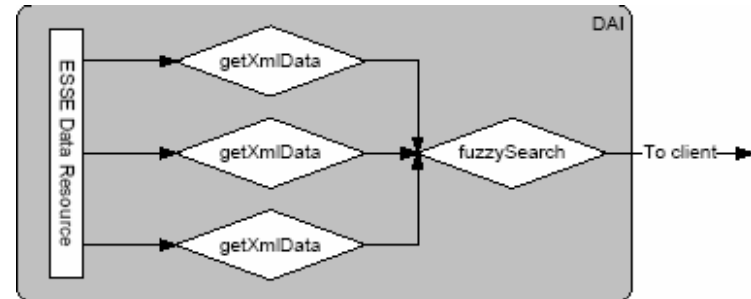


Pros for scientific applications:

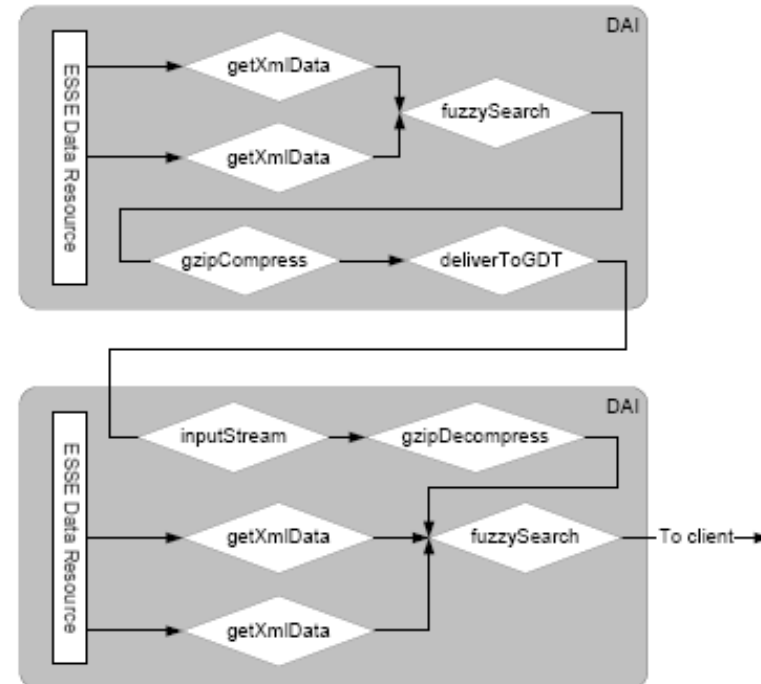
- Can be run both in GRID (WSRF, OMII) and pure web services container (Tomcat Axis)
- Data requests using XML allows data processing in heterogeneous environments
- Can be extended to access different types of data sources using activities and data resources

# Data flow management by OGSA-DAI

OGSA-DAI query from single data source



OGSA-DAI query from distributed data sources





# ESSE system components inside OGSA-DAI container

Component	Description
EsseDataResource	Represents environmental database
GetMetadataActivity	Query activity. Returns the description of the data maintained by the EsseDataResource.
GetXmlDataActivity	Query activity. Returns one or several time series from the EsseDataResource.
GetNetcdfDataActivity	Query activity. Serializes a data subset into a NetCDF file and returns an URL to that file.
FuzzySearchActivity	Transformation activity. Receives one or more time series from GetXmlData and returns fuzzy membership function values.

# Activities for data export

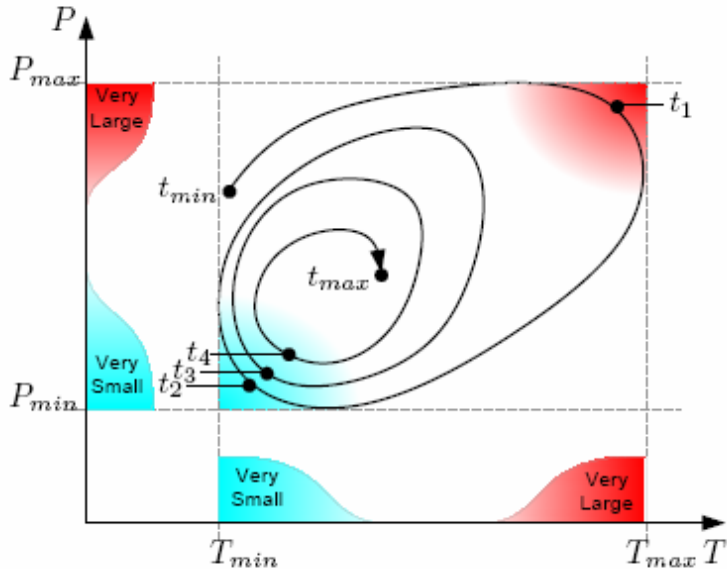
- XML output stream
  - We have plugin for NASA World Wind to visualize XML-formatted data
  - Can easily be transformed using XSLT to web page or another XML document, e.g. MS Excel
  - Can be used as input for ESSE fuzzy logic search engine
- NetCDF binary data file
  - Standard for scientific data storage in files
  - There are several visualization programs for NetCDF
  - Compatible with Unidata Common Data Model standard

# How to interpret a question of a scientist?

1. Introduce the notion of an Environmental Scenario (ES) as a basic building block for scientific question
2. Interpret ES as a fuzzy query expression
  - a. Each basic condition in a ES translates into membership function of a fuzzy set, a term in a resulting expression
  - b. An expression is built using traditional fuzzy logic operations plus “time shift” operator
3. Query terms are evaluated at individual data sources
4. The ESSE engine collects the data and performs fuzzy query operation.

The ESSE engine is built as a Web Service. This enables cascading queries, but raises new research challenges, e.g. optimization of query execution.

# Environmental scenario



Time series as a trajectory in the two-dimensional phase space (P-pressure, T-temperature)

**State  $S_1$**  corresponding to the red (upper-right) region is the fuzzy expression:

$$S_1 = (\text{VeryLarge } P) \text{ and } (\text{VeryLarge } T)$$

**State  $S_2$**  corresponding to the cyan (lower-left) region is:

$$S_2 = (\text{VerySmall } P) \text{ and } (\text{VerySmall } T)$$

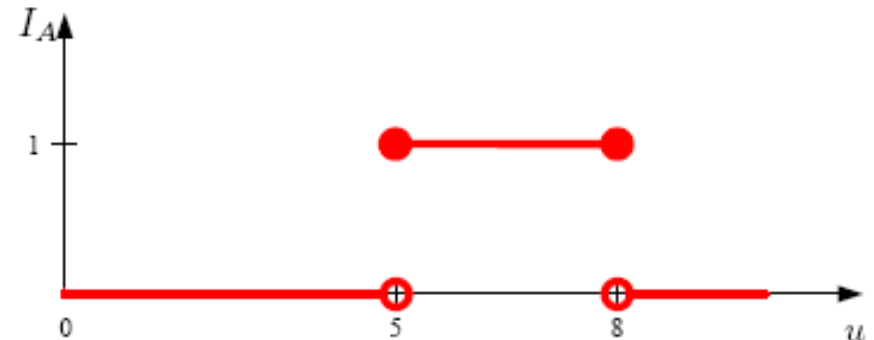
Combining the descriptions of the states with the **time shift operator**  $\text{shift}_{dt}$ , we can write the following symbolic expression for the **Environmental Scenario**

***“very low temperature and pressure after very high temperature and pressure”***:

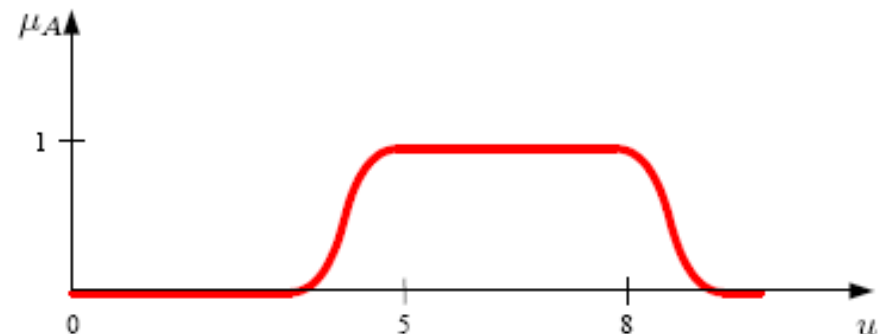
$$(\text{shift}_{dt=1} S_1) \text{ and } S_2$$

# Classical and fuzzy sets

Indicator function  $I_A(u)$  for the classical set  $A = \{x | 5 \leq x \leq 8\}$

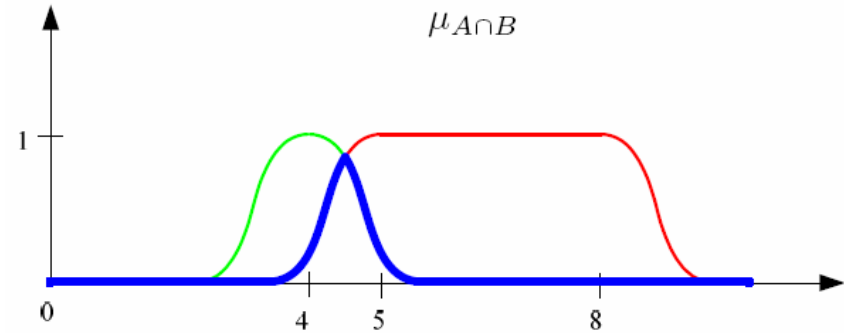


Fuzzy membership function  $\mu_A(u)$  for the set  $A = [5, 8]$

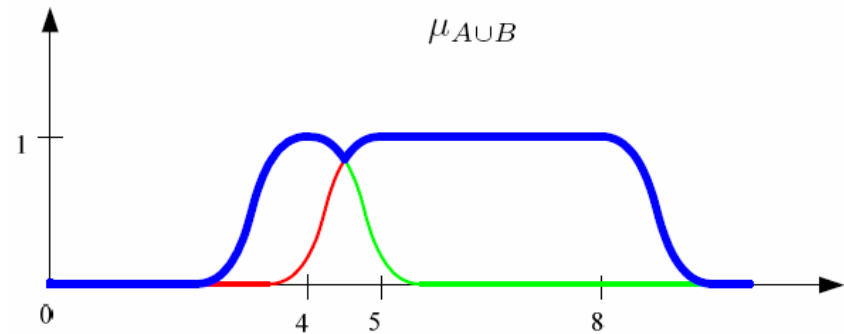


# Fuzzy logic operations

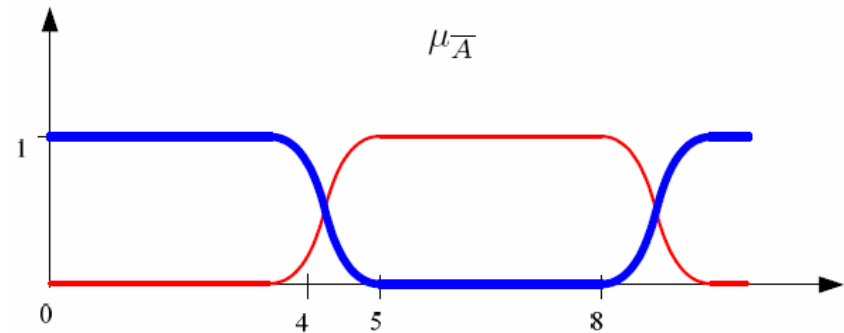
Intersection  
Fuzzy T-norm



Union  
Fuzzy T-conorm

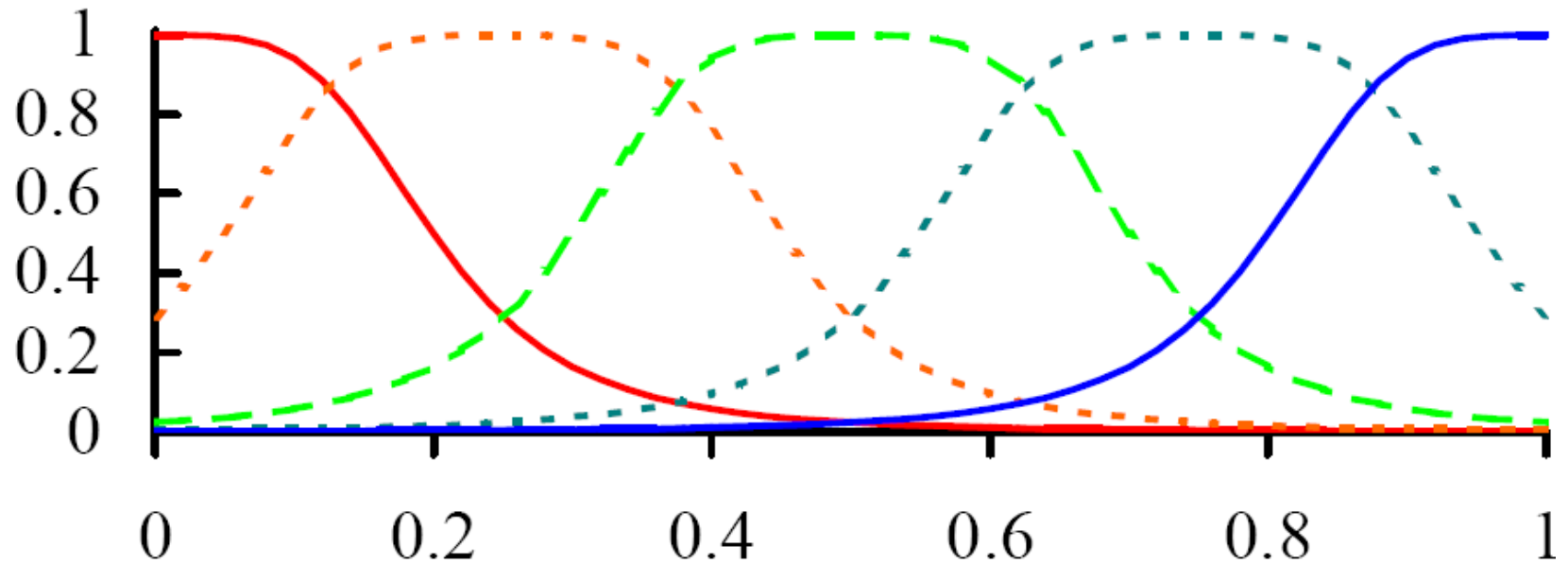


Logical not  
Fuzzy complement

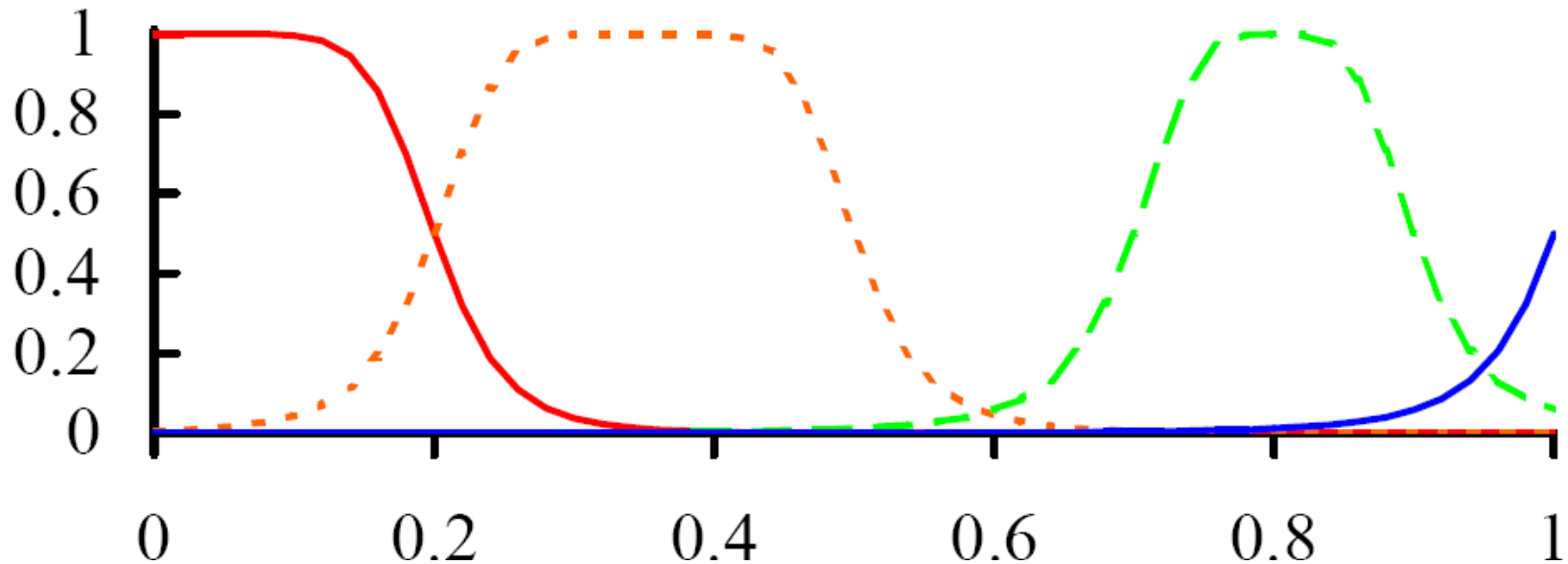


# Fuzzy logic predicates: “linguistic terms”

— Very small    - - - Small    - - - Average  
- - - Large    — Very Large



# Fuzzy logic predicates: “numerical terms”

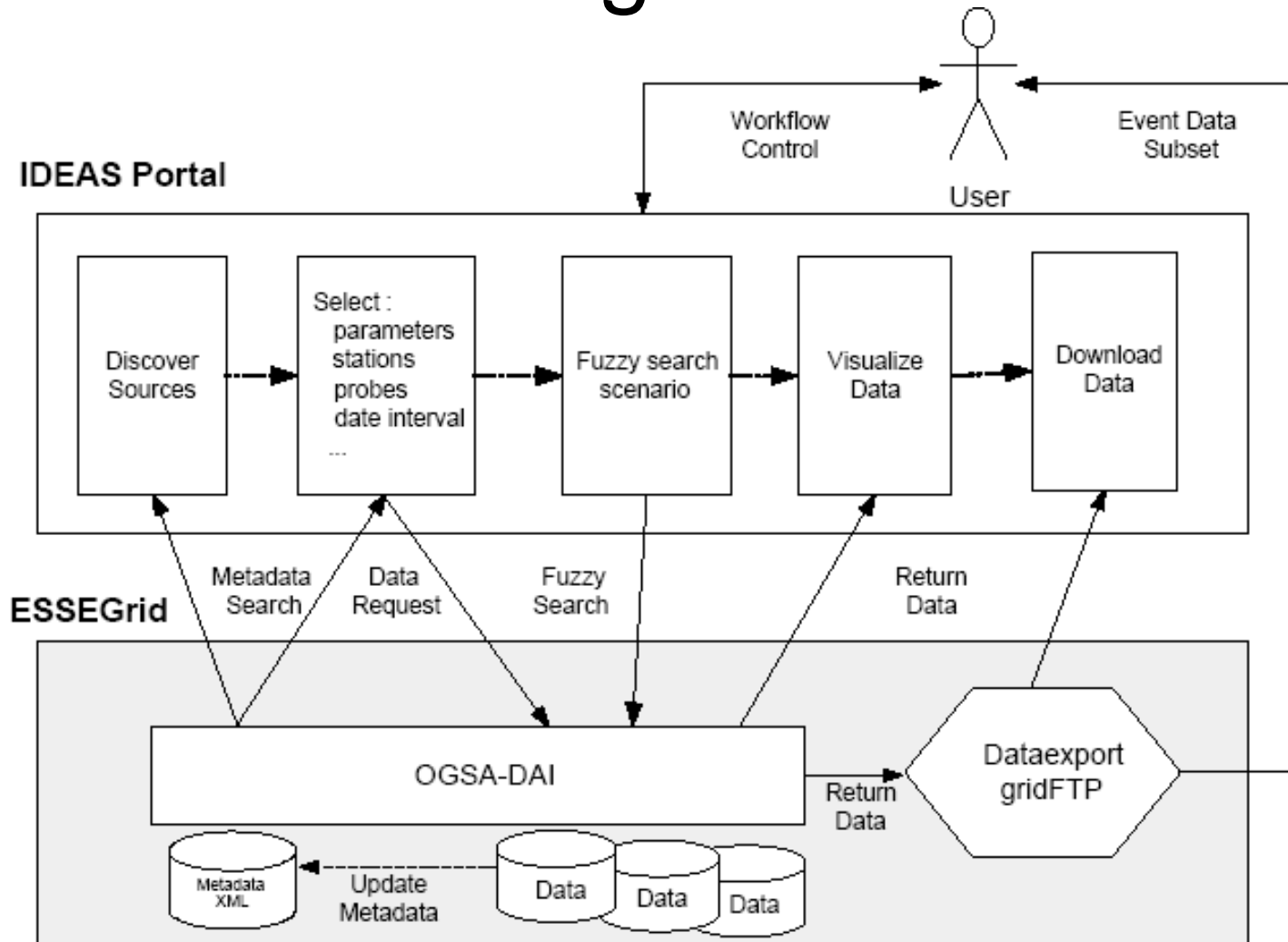




# How to synthesize and present results of a distributed query?

- Environmental Scenario search result is a scored list of candidate events. “Score” represents the “likeliness” of each event in a numerical form
- The result page provides links to visualization and data export pages
- Each event can be viewed as
  - time series
  - dynamic 5D volume
  - WorldWind color map on Earth surface
  - satellite images animation
- Data subset for each event can be exported in XML and NetCDF formats

# Web portal workflow using ESSE engine



# Web portal use case

In the following example we will search for a E-W atmospheric front near Moscow described by three parameters “air pressure”, “E-W wind speed” (Uwind) and “N-S wind speed” (V wind) with subsequent fuzzy states:

1. (Small pressure) and (Large V-wind-speed)
2. (Large pressure) and (Small U-wind speed)  
and (Small V-wind-speed).

# Step 1. Select data source

The user logs in to the IDEAS portal and receives a list of the currently available (distributed) data sources. For each data source the list has abridged metadata like name, short description, spatial and temporal coverage, parameters list and link to full metadata description. The user selects environmental data source based on the short description or by metadata keyword search (e.g. NCEP/NCAR Reanalysis).

**IDEAS** Integrated of Distributed Environmental Archives System

powered by ESSE

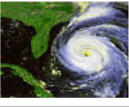
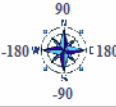
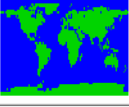

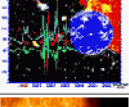

Login | **Databases** | Mining | Visualization | Documentation | Help

Select Database | ROI and Probes | Parameters

Login:  dimm  
Databases:

Parameters:  3 set  
Probes:  1 set  
ROI:

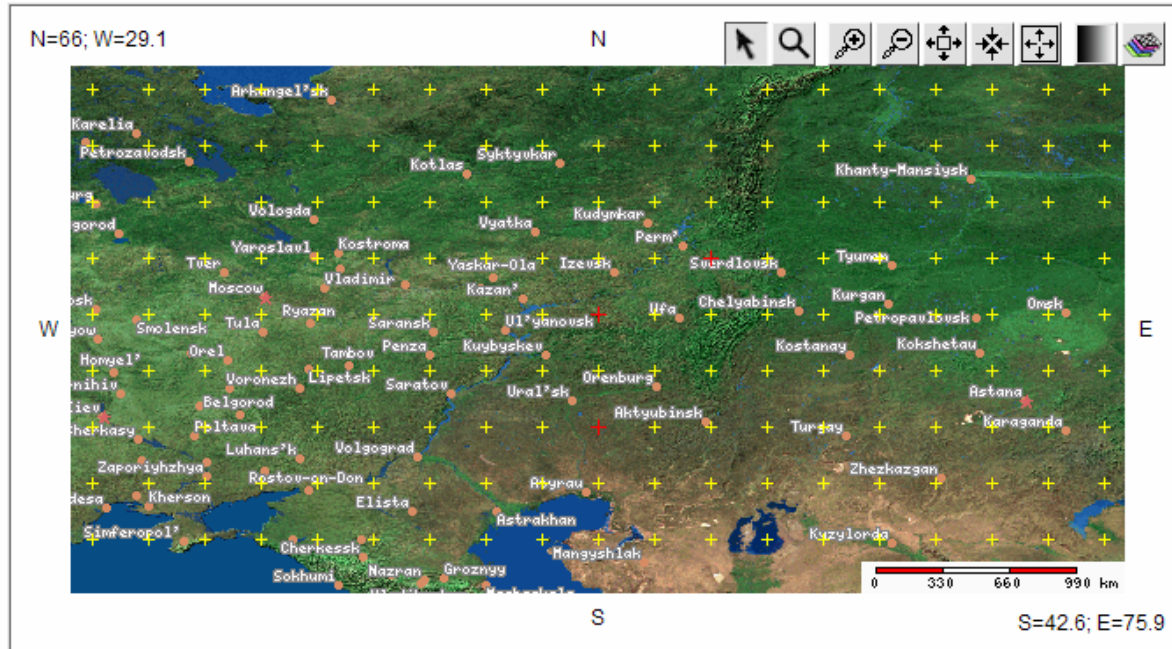
Scenario states:

Map	Database	Region	Time Span	Plot & Mine
	<b>NWS Weather Forecast [weather]</b> <ul style="list-style-type: none"><li><a href="#">View Database Description</a></li><li><a href="#">View Parameter List</a></li></ul>		January 01, 2005 thru December 31, 2006	<input type="checkbox"/>
	<b>NCEP/NCAR Reanalysis [ncep25]</b> <ul style="list-style-type: none"><li><a href="#">View Database Description</a></li><li><a href="#">View Parameter List</a></li></ul>		January 01, 1949 thru December 31, 2006	<input checked="" type="checkbox"/>
	<b>SPIDR Geomagnetic Variations (hourly) [spidr_geom_hr]</b> <ul style="list-style-type: none"><li><a href="#">View Database Description</a></li><li><a href="#">View Parameter List</a></li></ul>	N/A	January 01, 1901 thru December 31, 2003	<input type="checkbox"/>
	<b>SPIDR indices [spidr_indices]</b> <ul style="list-style-type: none"><li><a href="#">View Database Description</a></li><li><a href="#">View Parameter List</a></li></ul>	N/A	January 01, 1932 thru May 01, 2006	<input type="checkbox"/>

Source Name   
Source Description   
Parameter Name

# Step 2. Select spatial location

The portal stores the data source selection on the server side in the persistent “data basket” and presents a GIS map with the spatial coverage of the data source. The user selects a set of “probes” (representing spatial locations of interest, e.g. Moscow) for the searching event.



Enter Probe:	Latitude <input type="text"/>	Longitude <input type="text"/>
Enter ROI:	South <input type="text"/>	North <input type="text"/>
	West <input type="text"/>	East <input type="text"/>

Set probes & ROI Clear selection

# Step 3. Select environmental parameters

IDEAS stores the selected set of "probes" and presents a list of all the environmental parameters available from the selected data source and a fuzzy constraints editor on the parameters values which represent the event. The user selects some of the environmental parameters and sets the fuzzy constraints on them for the searching event (e.g. low pressure, high V-wind speed).

**Fuzzy Search Criteria**

**Fuzzy Membership Function**

		Linguistic					Numeric					Limits		Query		
		Very small	Small	Average	Large	Very large	≠	≈	Range	∩	Any value	Threshold1	Threshold2	Importance		
ncep25	<a href="#">WndUComp</a>		10	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="text" value="0.0"/>	<input type="text" value="0.0"/>	<input type="text" value="1.0"/>
ncep25	<a href="#">WndVComp</a>		10	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="text" value="0.0"/>	<input type="text" value="0.0"/>	<input type="text" value="1.0"/>
ncep25	<a href="#">PresMSL</a>		surface	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="text" value="0.0"/>	<input type="text" value="0.0"/>	<input type="text" value="1.0"/>

Update weather scenario state
Undo changes

# Step 4. Edit environmental scenario

Multiple subsequent environment states can be grouped to form the actual environmental scenario. For example, we need to define the two different states mentioned above. Adding and removing fuzzy states is done via a Web-form. ESSE stores the searching environment states and sends them to the fuzzy search web-service in the XML format.

**Temporal Extent**

Seasonal time intervals

Date Range: 19490101 to 20051231

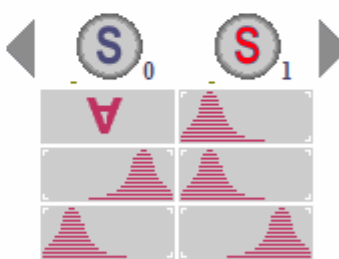
Date from, inclusive (year month day): 1950 ▾ Jan ▾ 1 ▾

Date to, inclusive (year month day): 2000 ▾ Dec ▾ 28 ▾

Time window: 1 day ▾

**Weather Scenario Fuzzy States**

Database	Parameter	Height	Units	
ncep25	<a href="#">WndUComp</a>	10	m/s	
ncep25	<a href="#">WndVComp</a>	10	m/s	
ncep25	<a href="#">PresMSL</a>	surface	Pa	

# Step 5. Search for events

The fuzzy search web-service collects data from the data source for the selected parameters and time interval, performs the data mining, and returns to the IDEAS web application a ranked list of candidate events with links to the event visualization and data export pages.

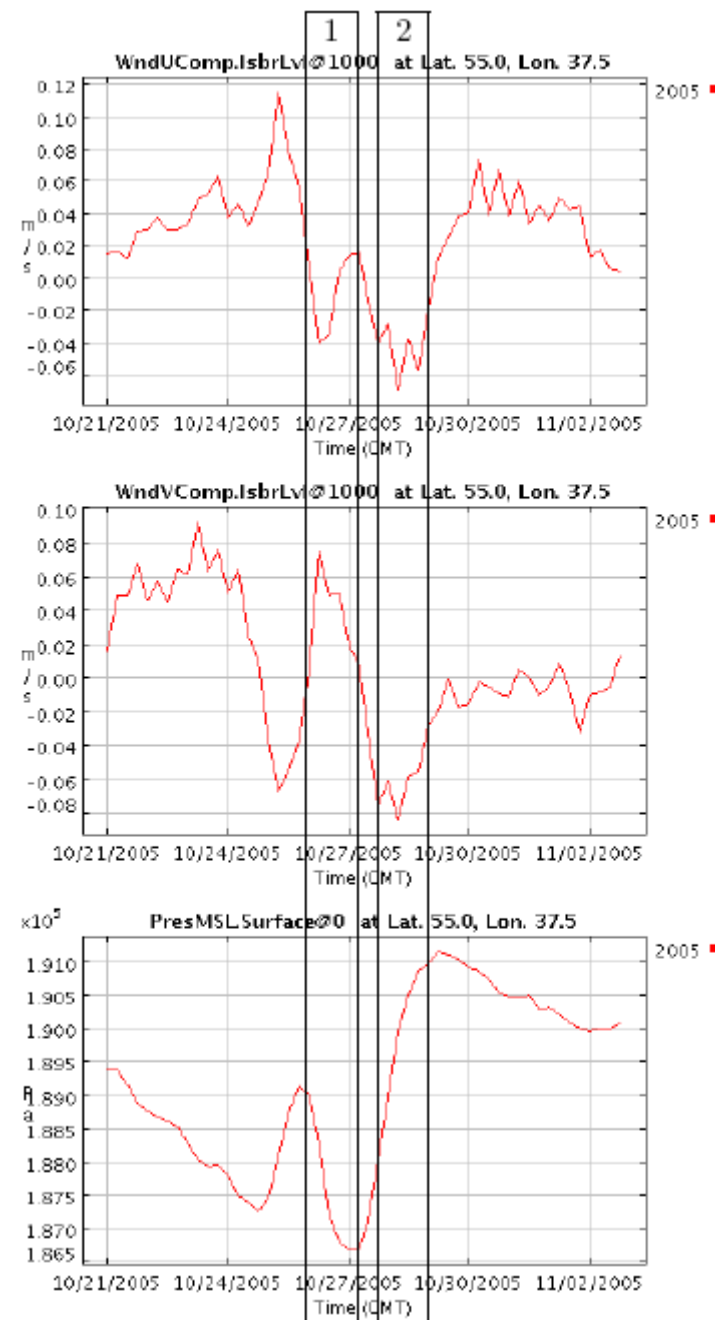
Data search took: 43.698 seconds.

Weather Scenario Search Results						
Rank	Score	Date	Time	Probes	ROI	Satellite
1	0.23	1950-08-13	11:00:00	<a href="#">Plot</a>	<a href="#">Vis5d</a>	<a href="#">DMSP</a>
2	0.20	1950-07-08	6:00:00	<a href="#">Plot</a>	<a href="#">Vis5d</a>	<a href="#">DMSP</a>
3	0.16	1950-06-08	11:00:00	<a href="#">Plot</a>	<a href="#">Vis5d</a>	<a href="#">DMSP</a>
4	0.16	1950-09-30	22:00:00	<a href="#">Plot</a>	<a href="#">Vis5d</a>	<a href="#">DMSP</a>
5	0.16	1950-04-12	23:00:00	<a href="#">Plot</a>	<a href="#">Vis5d</a>	<a href="#">DMSP</a>
6	0.15	1950-09-12	14:00:00	<a href="#">Plot</a>	<a href="#">Vis5d</a>	<a href="#">DMSP</a>
7	0.13	1950-02-06	8:00:00	<a href="#">Plot</a>	<a href="#">Vis5d</a>	<a href="#">DMSP</a>
8	0.12	1950-06-12	16:00:00	<a href="#">Plot</a>	<a href="#">Vis5d</a>	<a href="#">DMSP</a>
9	0.12	1950-05-26	2:00:00	<a href="#">Plot</a>	<a href="#">Vis5d</a>	<a href="#">DMSP</a>
10	0.11	1950-07-12	22:00:00	<a href="#">Plot</a>	<a href="#">Vis5d</a>	<a href="#">DMSP</a>

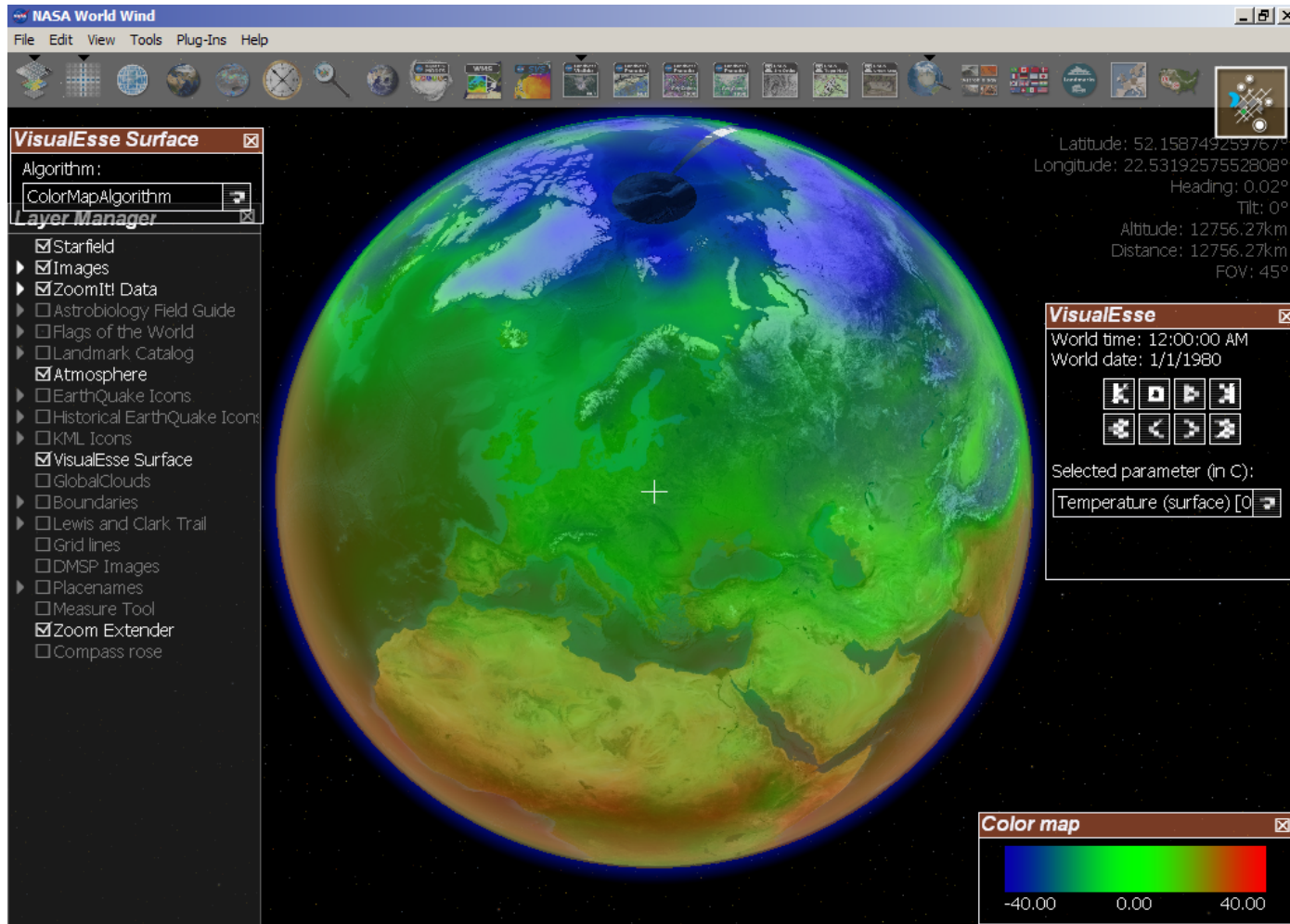


# Step 6. Visualize event

The user visualizes interesting events and requests the event-related subset of the data for download from the data source in the preferred scientific format (XML, NetCDF, CSV table). Currently there are four visualization types available: time series, animated volume rendering using Vis5D, DMSP satellite images and NASA WorldWind visualization.



# Step 7. XML-formatted data with NASA WorldWind



# Step 8. Event view from DMSP satellite

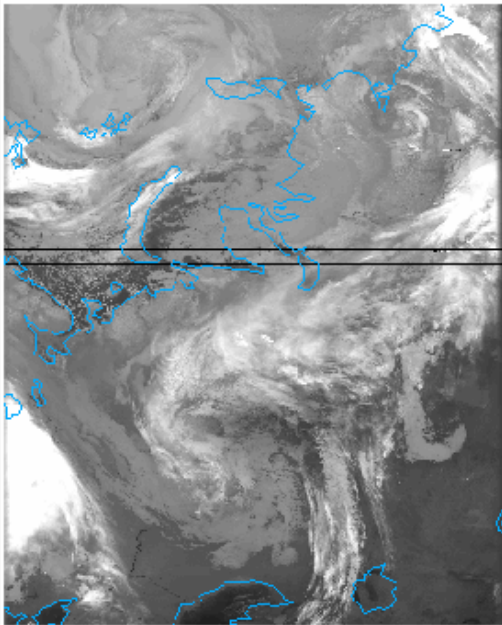
**Time Control**  
Current time (GMT): 2005-10-26 03:28:13, satellite: F13, day

Time interval: 2005102600 - 2005102723  
Selected location: Lat=55.0' deg Lon=37.5' deg  
Images for the location: 47

**Orbit Control**

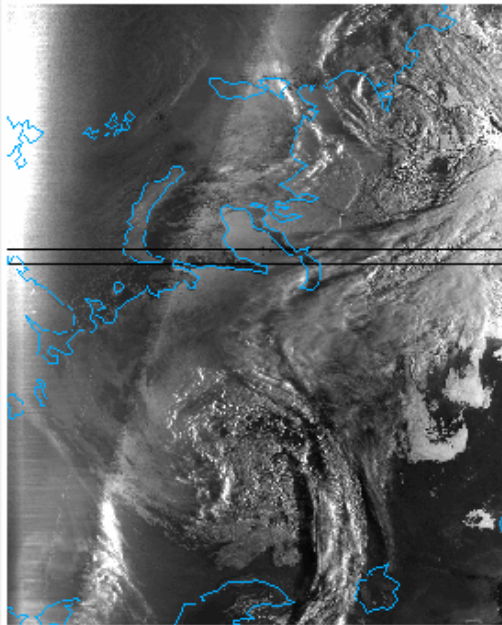
Orbit N+1, Previous 1/8, Reset, Next 1/8, Orbit N-1

**DMSP Infrared Channel**



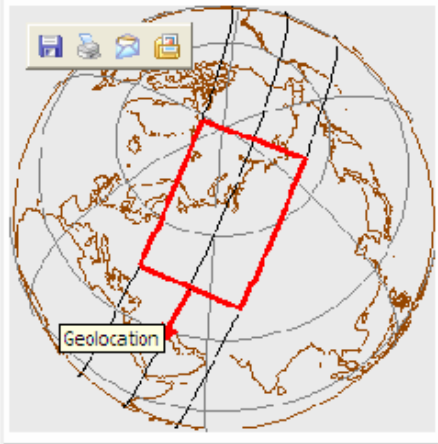
F13200510260302\_3\_ols tir.png

**DMSP Visible Channel**



F13200510260302\_3\_ols vis.png

**1/8th Orbit**



Geolocation

# CLASS: Comprehensive Large Array-data Stewardship System. Portal prototype.

Supported data:

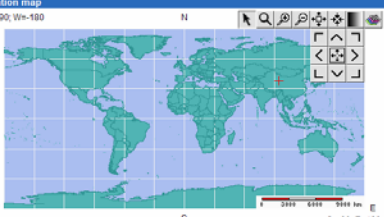
- Time series:
  - NCEP/NCAR weather reanalysis (ESSE)
  - Geomagnetic indices database – SPIDR
  - Ionospheric data – SPIDR
  - Sea surface temperature – NGDC NOAA
- Satellite images
  - DMSP
  - MODIS
  - CLASS (AVHRR)

The screenshot displays the CLASS portal interface. At the top, there is a navigation bar with the NOAA logo and the text "NOAA Satellite and Information Service" and "National Environmental Satellite, Data, and Information Service (NESDIS)". A search bar is located on the right side of the header. Below the header, the main content area is divided into several sections. On the left, there is a sidebar menu with categories like "Granules" (listing DMSP, MODIS, AVHRR), "Time Series" (listing NCEP Reanalysis, Geomagnetic indices, Ionosphere, Sea surface temperature), "Login" (with a login form), "Support" (with Register, Help, Services), and "User items" (with Shopping cart, User preferences). The main content area features an "Introduction to CLASS" section, which includes a paragraph describing the system's purpose and a "CLASS" logo. Below this, there is a "Search to Product" section with three search boxes for Keywords, Location (Lat Lon), and Time (yyyy-mm-ddThh:mm:ss), each with a "Search" button. On the right side, there is a "Last News" section with two news items, each titled "NOAA-16 Data Declared Operational (1-24-01)" and providing details about the data's availability.

# Fuzzy search for CLASS

**Detailed CLASS data request**

**Time window**  
 Date from: 2005 Dec 24 0 0  
 Date to: 2005 Dec 31 23 59

**Location map**  
 N=90, W=-180  
  
 W S E  
 S=-90, E=180

**Product Description**  
 NCEP/NCAR Reanalysis  
 The NCEP/NCAR Reanalysis 1 project is using a state-of-the-art analysis/forecast system to perform data assimilation using past data from 1948 to the present. A large subset of it is available from CDC in its original 4 times daily as daily averages. However, the data from 1: a 1hr different, in the regular (non-Gaussian) data. That data was done at 6 times daily in it because the inputs available in that era were 3C, 9C, 12C, and 21C, whereas the 4x daily C been available at 0C, 0C, 12C, and 18C. These times were forecasted and the combined reanalysis era is 6x daily. The local ingestion period only the 0C, 0C, 12C, and 18C forecasted values thus only those were used to make the daily 1 and monthly means here.

**Fuzzy Search Criteria**

Fuzzy Membership Function	Linguistic	Numeric	Limits	Query
Temperature[C]	Very small, Small, Average, Large, Very large	is, Range, Any value, Threshold1	Threshold1, Threshold2	Expands
Total Cloud Cover[%]			0.0, 1.0	1.0
Precipitation Rate[kg/m/m/s]			0.0, 1.0	1.0

Action: Event search

CLASS portal can filter satellite orbits database search for given location based on the fuzzy event definition such as Low Cloud Coverage (cloud free orbits) or magnetic storm (Aurora images).

**CLASS**  
[+ home](#)

NOAA > NESDIS > NGDC > STP > CLASS  
 Comprehensive Large Array-data Stewardship System

NOAA Satellite and Information Service  
 National Environmental Satellite, Data, and Information Service (NESDIS)

**Granules**  
 DMSP  
 MODIS  
 AVHRR

**Time Series**  
 NCEP Reanalysis  
 Geomagnetic indices  
 Ionosphere  
 Sea surface temperature

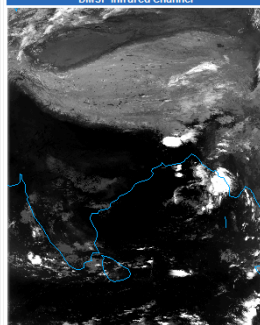
Login

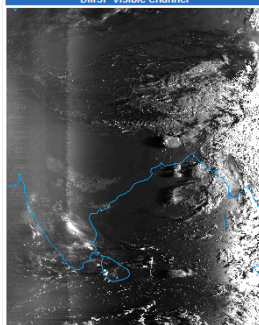
Support  
 Register  
 Help  
 Services

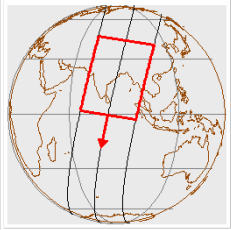
User items  
 Shopping cart  
 User preferences

**Time Control**  
 Current time (GMT): F14200610112341.4, satellite: F14, day

**Orbit Control**

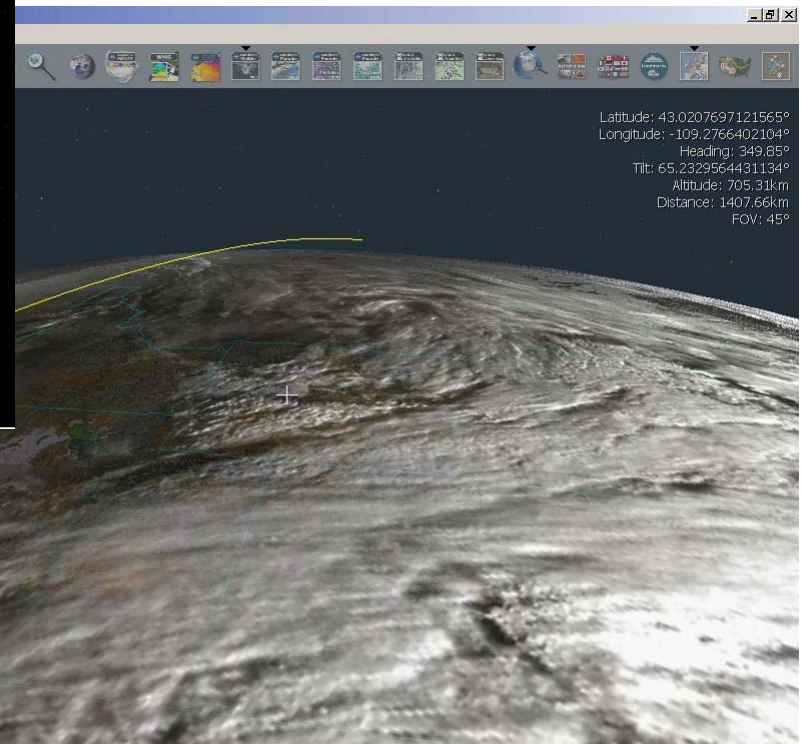
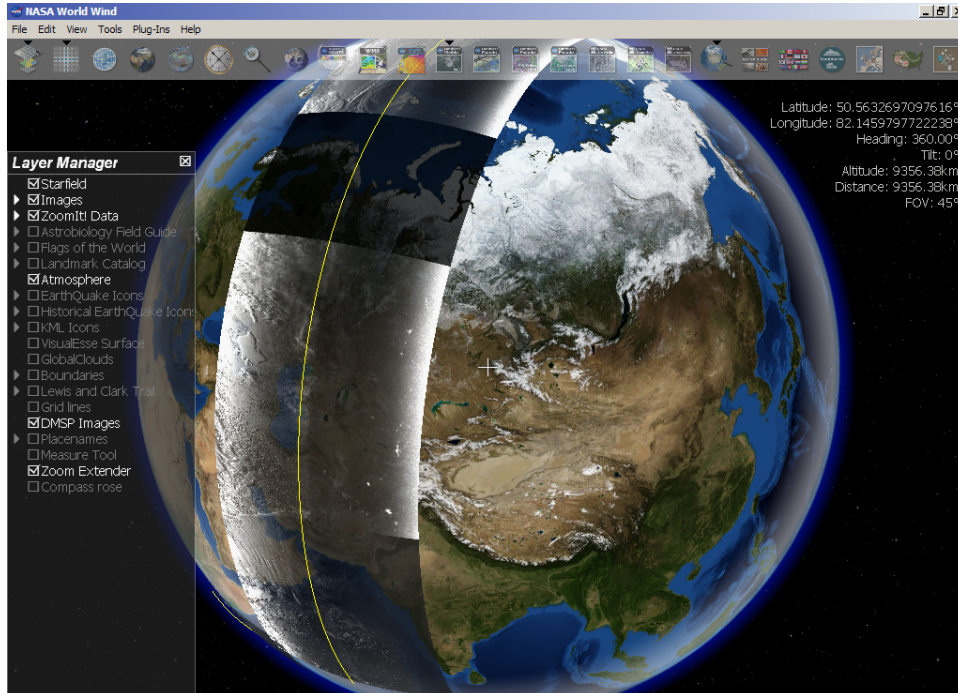
**DMSP Infrared Channel**  
  
 F14200610112341.4\_ols.tr.png

**DMSP Visible Channel**  
  
 F14200610112341.4\_ols.vis.png

**1/8th Orbit**  




# DMSP orbits visualization with NASA world wind



# Fuzzy data mining is used by:



ESSE <http://esse.wdcb.ru/>

Environmental Scenario Search Engine

The main idea behind ESSE is a flexible, efficient and easy to use search engine for data mining in environmental data archives.

The project is supported by Microsoft Research, Cambridge, and NOAA



CLASS <http://spidr.noaa.gov/class/>

Comprehensive Large Array-data Stewardship System

CLASS is NOAA's premier on-line facility for the distribution of NOAA and US Department of Defense (DoD) Polar-orbiting Operational Environmental Satellite (POES) data and derived data products



DEGREE <http://degree.ipgp.jussieu.fr/>

Dissemination and Exploitation of GRids in Earth science

The project aims to promote the GRID culture within the different areas of ES and to widen the use of GRID infrastructure as platform for e-collaboration in the science and industrial sectors and for select thematic areas which may immediately benefit from it

# Thank you

<http://esse.wdcb.ru>

<http://spidrd.ngdc.noaa.gov/class>

[esse@wdcb.ru](mailto:esse@wdcb.ru)