Towards a cloud-based computing and analysis framework to process environmental science big data

Eleonora Luppi, Sebastiano Fabio Schifano, Luca Tomassetti

University of Ferrara, Italy

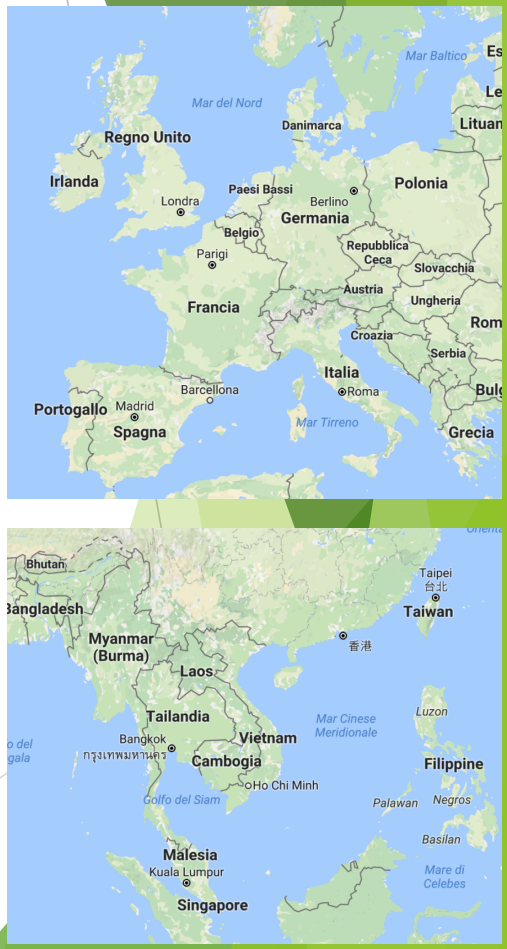
Introduction

- Environmental sciences use data coming from several sources:
  - satellites
  - large network of sensors installed on the ground or sea-floating stations
  - devices installed on balloons or aircrafts
- These networks produce a big amount of data that needs to be appropriately processed and analyzed to extract information useful for scientists to investigate natural phenomenas
- Needs:
  - to collect and store huge amount of data together with space and time information
  - large and powerful computing resources to run analysis and visualization codes

TORUS Project

Toward Open Resources Using Services

- Interdisciplinary EU - ERASMUS+ Capacity Building - **TORUS** project, which includes Europe's and South East Asia's partners with a strong expertise in distributed computing and earth and environmental sciences.
- **TORUS** project aims at making available to environmental scientists a cloud based computing and analysis framework to manage and process big-data:
  - ability to access clouds to virtualize the computing resources, and knowledge to use software tools to process and analyze data coming from the different sources
  - data correlation with time and space meta-data information and data storage
  - high-level data presentation to facilitate management and analysis by user scientists
  - investigation of high-performance computing integration to boost tasks, also using recent accelerators like GP-GPUs or many-core processors



TowardOpenResourcesUsingServices

Co-funded by the Erasmus+ Programme of the European Union

TORUS Project

- Partners:

- Regular Workshops:
  - Hanoi (Jan, 2016)
  - Ferrara (Jun, 2016)
  - Pathumthani (Nov, 2016)
  - Brussel (Mar, 2017)
  - Ho Chi Min (Sep, 2017)
  - Wailalak Univ. (2018)
  - Pau (2018)

Co-funded by the Erasmus+ Programme of the European Union

TORUS Project Goals

- **Develop research on cloud computing in the environmental sciences and promote its education in the countries of South East Asian partners.**
- Installation of two computation mini-clusters with private cloud:
  - VNU – Hanoi
  - AIT – Pathumthani

  - Dual-socket CPUs (>10 cores each)
  - 64GB of RAM per socket
  - 2x10Gbits network
  - ~100TB storage server with SSD cache
  - Linux based (Debian) OS

- Setup will be finalized in H2 2017

TORUS Project

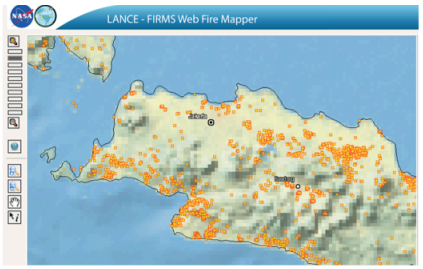
- Several applications in Earth and environmental sciences, geography, satellite image processing are the main focus of the project partners:
  - AIT: Air Pollution Modeling Applications in Thailand
  - VNU: Air Pollution Mapping from Space in Vietnam
  - VUB: Water Resources Management
  - Toulouse: Statistical approach to geographic applications

AIT - Air Pollution Modeling Applications

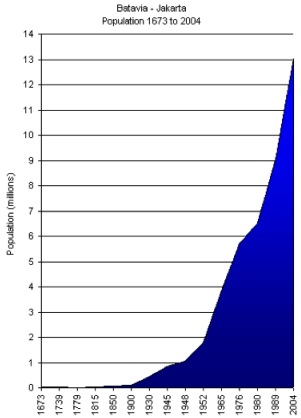
- Dr. D. A. Permadi, Prof. N. T. Kim Oanh
- Asian Institute of Technology, Pathumthani, Thailand

AIT - Air Pollution Modeling Applications

Rapid development of the city (Parasati, 2011)



Biomass open burning practices

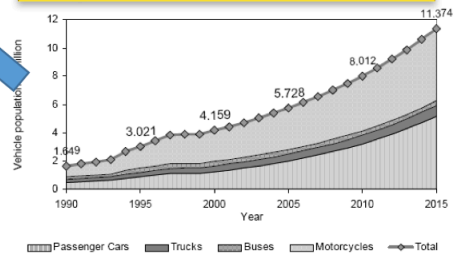


High population growth

Air Pollution Problems:
- Health impacts
- Ecosystem impacts

Urgent calls for mitigation

High rate of motorization



TowardOpenResourcesUsingServices

Co-funded by the Erasmus+ Programme of the European Union

AIT - Air Pollution Modeling Applications

- Environment effects are product of complex dynamic system driven by multiple processes (e.g. main processes determining air pollutant dispersion)
  - Atmospheric transport by mean wind field
  - Atmospheric turbulent diffusion
  - Atmospheric chemical and photochemical reactions
  - Interactions between surface (sea, land) and atmosphere
  - Wet and dry removal process
- Modeling tool used to integrate these processes in a systematic approach to assess impacts of different scenarios on environment (causal links)
- Hindcast, nowcast, and forecast are possible

AIT - Air Pollution Modeling Applications

Dispersion Model

based on a set of analytical or numerical algorithms (mathematical equations) describing physical, chemical aspects of air pollution to estimate ambient concentrations

Receptor Model

These models are observational techniques which use the chemical and physical characteristics of gases and particles measured at source and receptor to both identify the presence of and to quantify source contributions to receptor concentrations

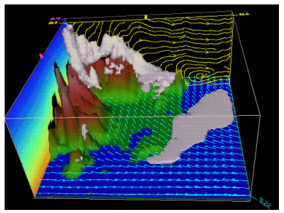
Statistical Model

Statistical models for predictions: time series, synoptic climatological models, etc

Source emissions

Concentration at receptors

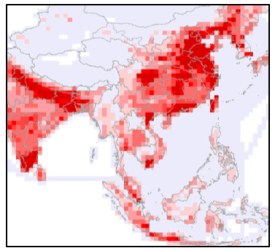
AIT - Air Pollution Modeling Applications



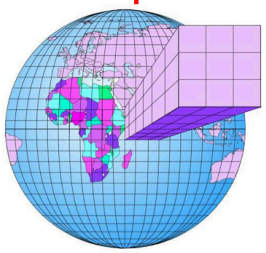
3D- Meteorology Data (output of meteorology model)

Atmospheric Concentration

Surface Deposition

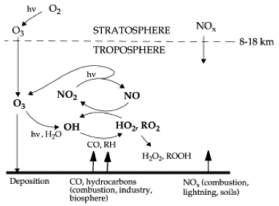


Emission data of SO_2, NOx and NH_3 as output of emission inventory

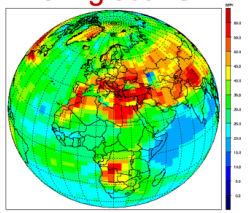


Chemistry Transport /Dispersion Model

Chemistry Mechanisms



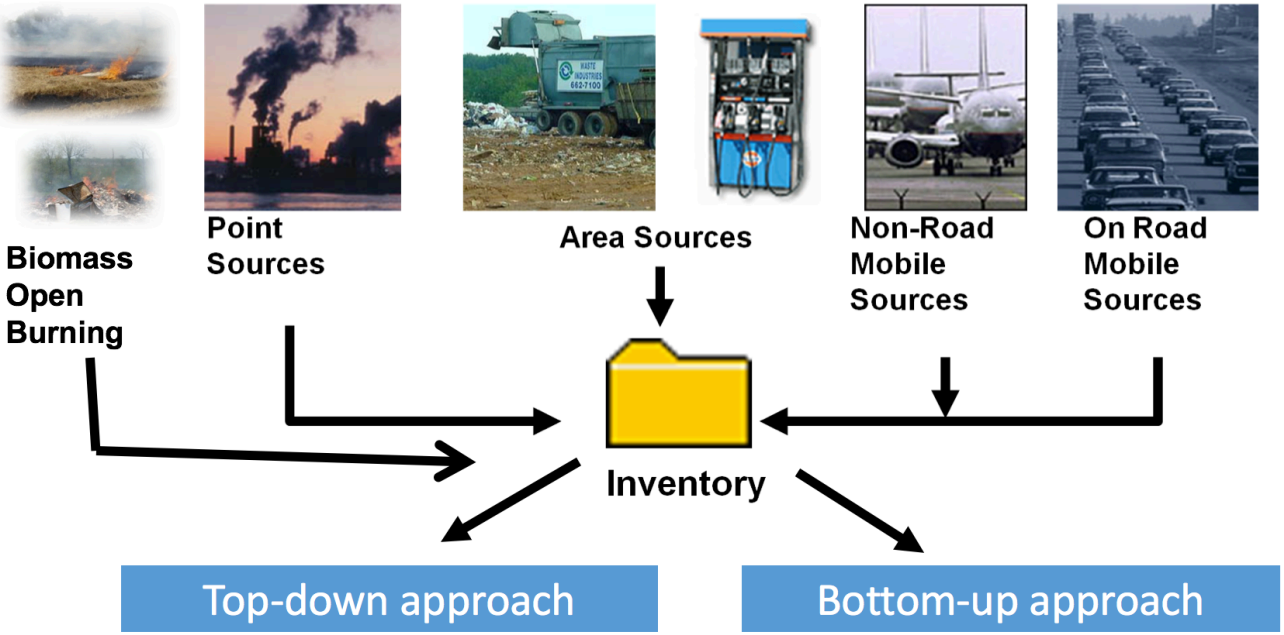
Boundary condition from global CTM



Land cover data

AIT - Air Pollution Modeling Applications

Emission Inventory - a comprehensive listing by sources of air pollutant emissions in a geographic area during a specific time period



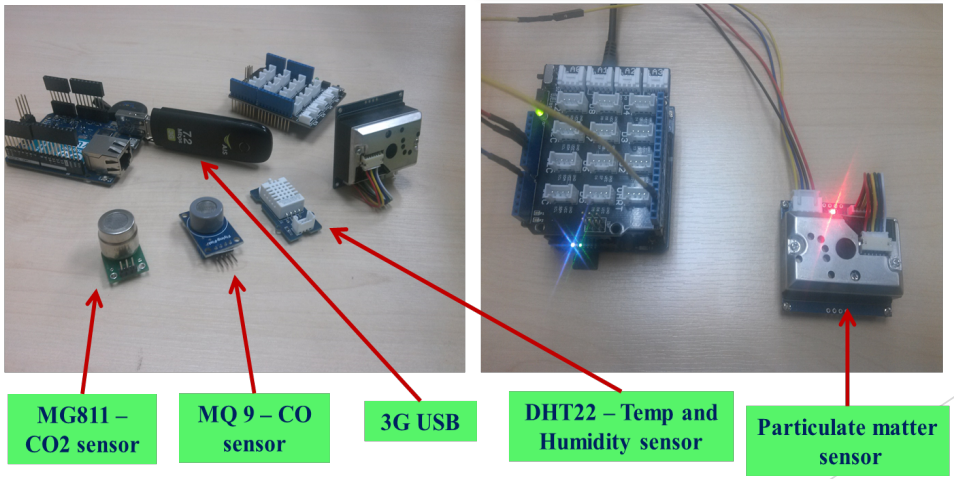
Co-funded by the Erasmus+ Programme of the European Union

AIT - Air Pollution Modeling Applications

- Air quality models require extensive data transfer and storage (input – output of meteorology and chemistry)
  - Satellite images and metadata from MODIS/VIIR S/LandSat/etc…, albedo, green fraction, land-use, USGS landcover, orography, soil type, and topography
  - The Emission Database for Global Atmospheric Research (EDGAR),
  - The Atmospheric Composition Change by the European Network of Excellence (ACCENT),
  - The Regional Emission inventory in ASia (REAS),
  - Global Fire Emission Database (GFED)
  - Inventory for: Ozone, NO_x, CO_2, SO_2, CO, N_2O, NH_3, Black-Carbon, Organic-Carbon, CH_4, $PM_{2.5}$, Total Particulate Matter, and Non-Methan Volatile Organic Compounds
- **High performance computing is important for model simulations**
- Integrated application for data visualization/dissemination through web-based interface can be developed using Cloud services

AIT - Air Pollution Modeling Applications

- Network of connected ground sensors
- PaaS for retrieval & visualization of collected data



TowardOpenResourcesUsingServices

Co-funded by the Erasmus+ Programme of the European Union

AIT - Air Pollution Modeling Applications Main Components

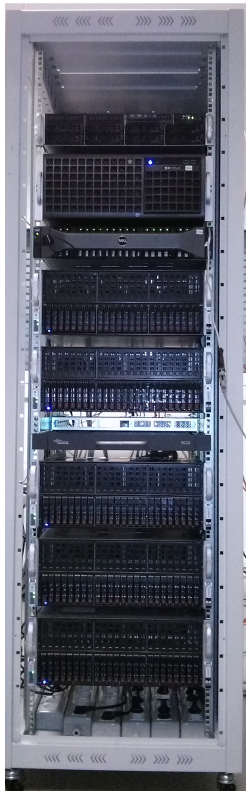
- **Atmospheric modeling system**
- Meteorological model (WRF: Weather Research and Forecasting)
  - Developed by National Center for Atmospheric Research (NCAR) and National Oceanic and Atmospheric Administration (NOAA): it's a supported *community model* with free and shared resources and distributed development.
  - 2 dynamical cores:
    - NMM (Nonhydrostatic Mesoscale Model) for atmospheric physics, real-time and forecast.
    - **ARW** (Advanced Research WRF) for global and regional climate, coupled-chemistry applications, and idealized simulations.
- Chemistry Transport Models (Chimere and CAMx)
  - Chimere is a multi-scale model primarily designed to produce **daily forecasts of ozone, aerosols and other pollutants** and make **long-term simulations** for emission control scenarios
  - Comprehensive Air quality Model with eXtensions (CAMx) is an open-source modeling system for multi-scale integrated assessment of gaseous and particulate air pollution.

TowardOpenResourcesUsingServices

Co-funded by the Erasmus+ Programme of the European Union

Test and prototyping

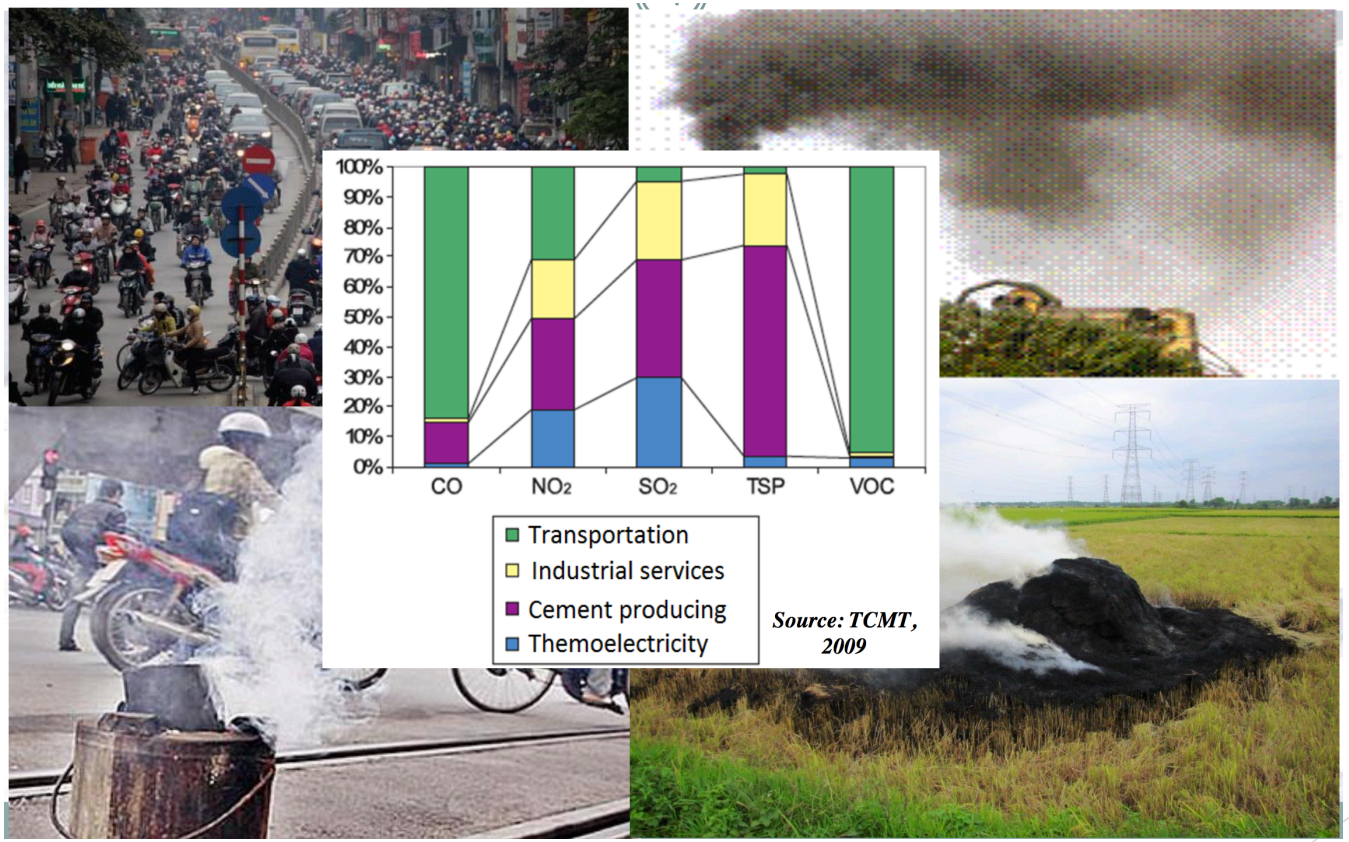
- Collaboration between Unife and AIT to early prototyping and optimization of WRF / air pollution modeling applications in a HPC cluster
- Use of the Ferrara's cluster
  - 5 nodes with 2 CPUs, 8 cores per CPU
  - 2 Infiniband FDR per node
  - 8 dual GPU Nvidia K80 per node
- Goal:
  - optimized run @AIT and @VNU clusters
  - future exploitation of GPU computing



VNU – Air Pollution Mapping from Space

- NGUYEN THI NHAT THANH, BUI QUANG HUNG, LE THANH HA, NGUYEN NAM HOANG, NGUYEN HAI CHAU, NGUYEN THANH THUY, PHAM VAN HA, LUU VIET HUNG, MAN DUC CHUC, PHAM NGOC HAI, PHAM HUU BANG, LE XUAN THANH PHAN VAN THANH, DO XUAN TU
- CENTER OF MULTIDISCIPLINARY INTEGRATED TECHNOLOGIES FOR FIELD MONITORING
  UNIVERSITY OF ENGINEERING AND TECHNOLOGY, VIETNAM NATIONAL UNIVERSITY HANOI

VNU – Air Pollution Mapping from Space



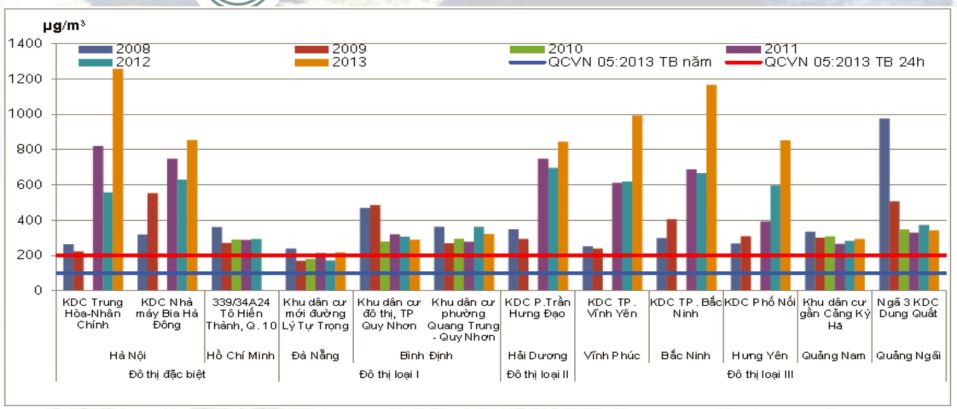
TSP: Total Suspended Particles
VOC: Volatile Organic Compounds

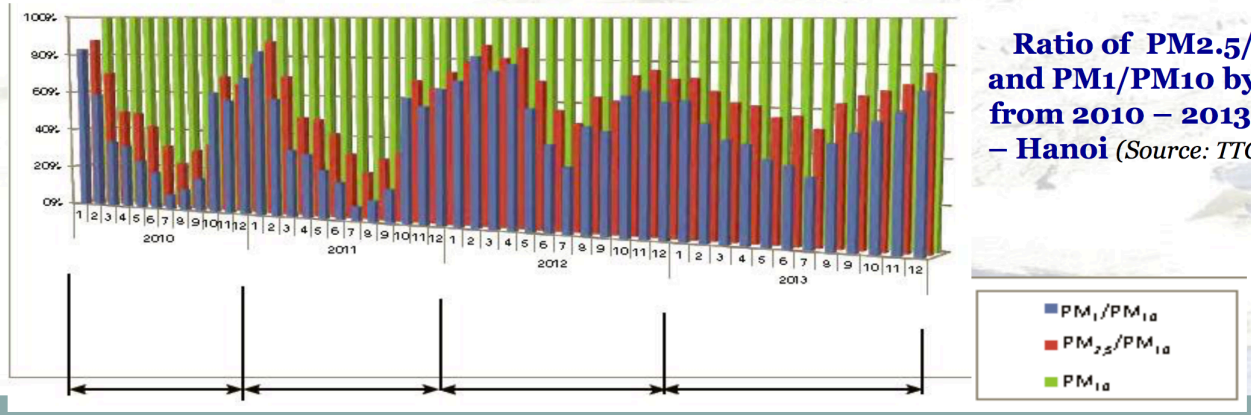
TowardOpenResourcesUsingServices

Co-funded by the Erasmus+ Programme of the European Union

VNU – Air Pollution Mapping from Space

Yearly TSP at cities in Vietnam from 2008 to 2013
(Source: TCMT, 2013)





Ratio of PM2.5/PM10 and PM1/PM10 by month from 2010 – 2013 at NVC – Hanoi *(Source: TTCMT, 2013)*

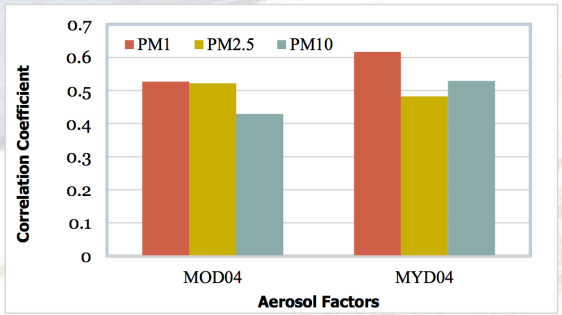
TowardOpenResourcesUsingServices

Co-funded by the Erasmus+ Programme of the European Union

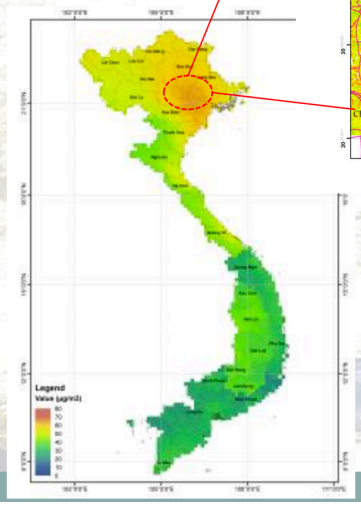
VNU – Air Pollution Mapping from Space

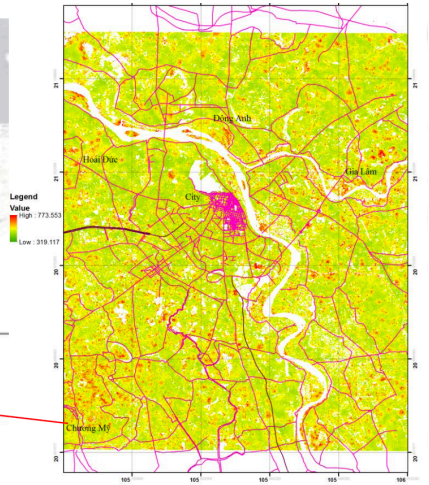
- Monitor PM based on satellite images

MODIS AOT vs. PM Correlation at
PhuTho, Hanoi, Hue, DaNang, 2010-2014



- Provide products at different spatial and temporal scales





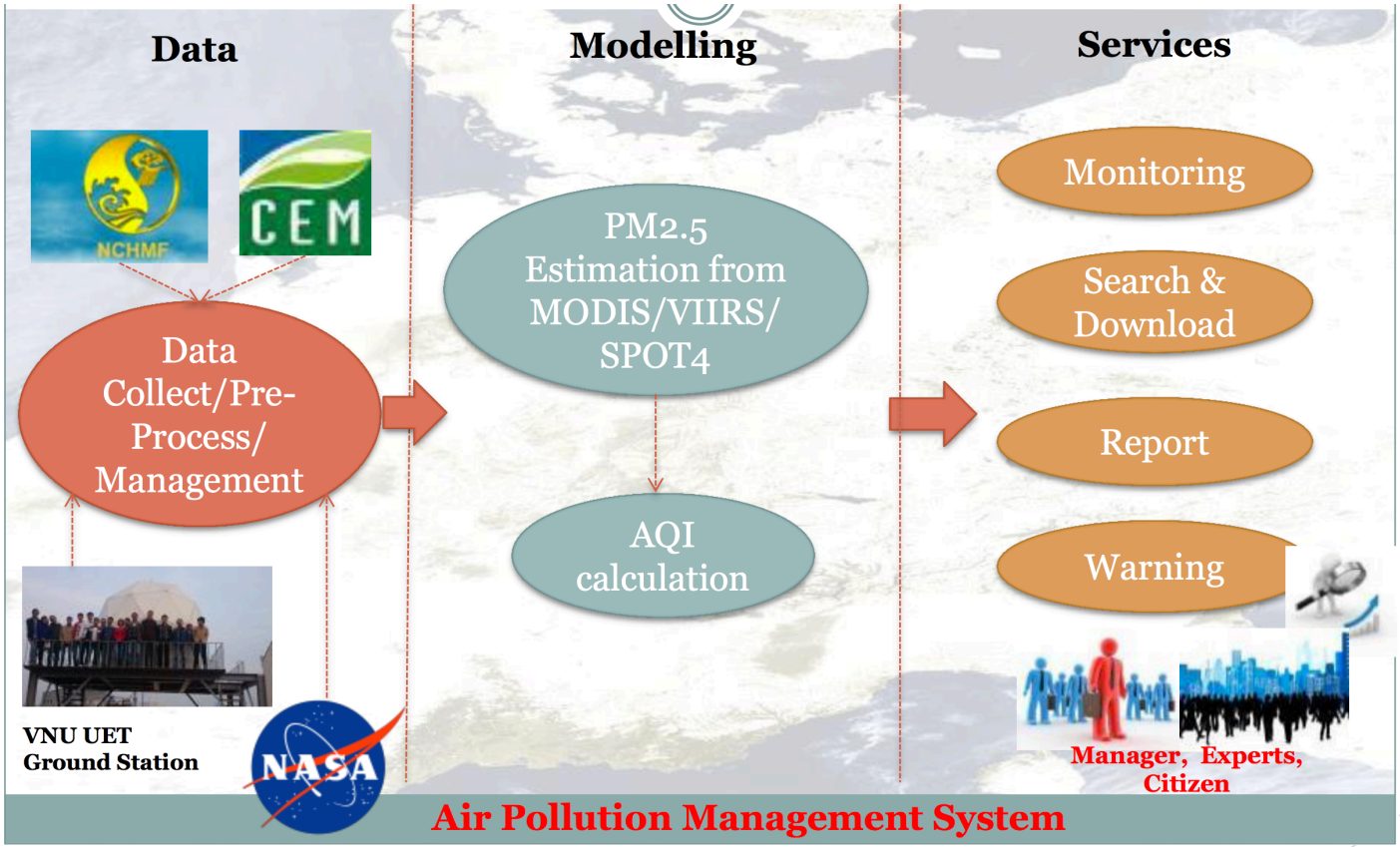
PM10 estimated from SPOT4 at May 4, 2012

PM2.5 Estimation from MODIS in Jan 24th, 2014

TowardOpenResourcesUsingServices

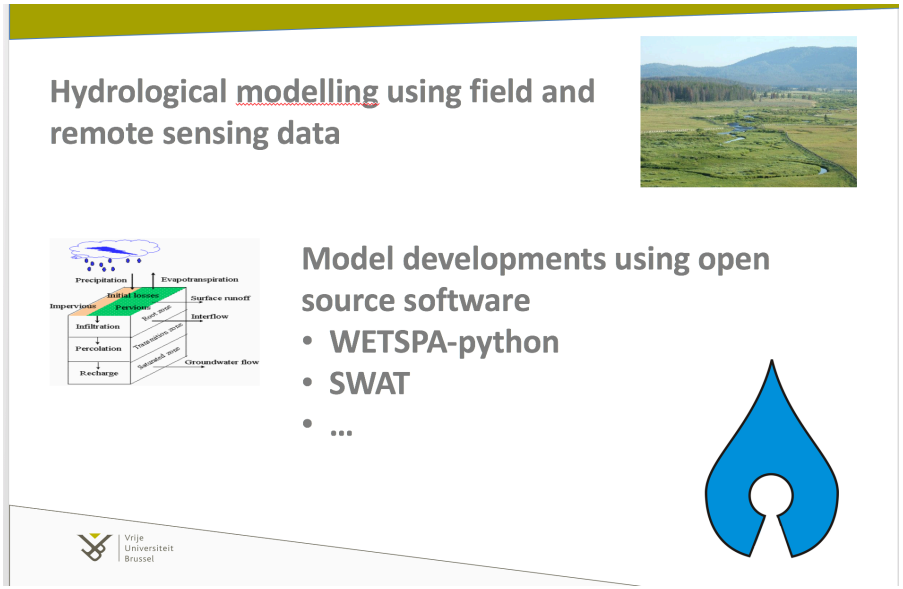
Co-funded by the Erasmus+ Programme of the European Union

VNU – Air Pollution Mapping from Space



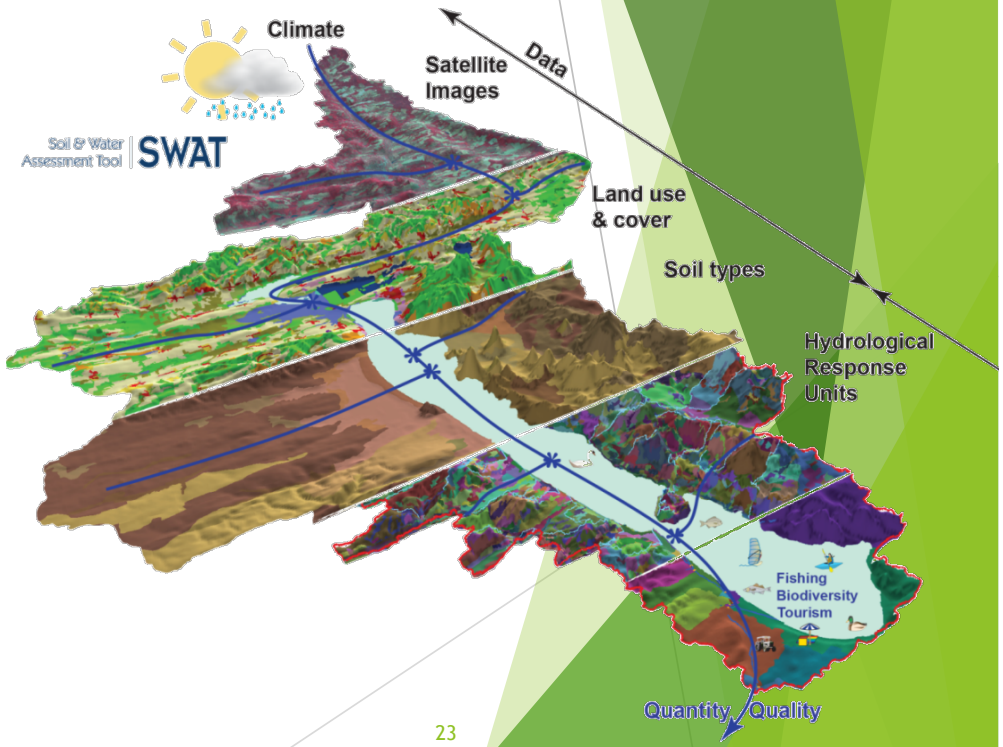
VBU - Water resources management

- Ann van Griensven, Hichem Sahli, Imeshi Weerasinghe
- Vrije Universiteit Brussel



VBU - Water resources management

- The Soil and Water Assessment Tool (SWAT) is a public domain model jointly developed by USDA Agricultural Research Service (USDA-ARS) and Texas A&M AgriLife Research, part of The Texas A&M University System.
- SWAT is a small watershed to river basin-scale model to simulate the quality and quantity of surface and ground water and predict the environmental impact of land use, land management practices, and climate change.
- SWAT is widely used in assessing soil erosion prevention and control, non-point source pollution control and regional management in watersheds.

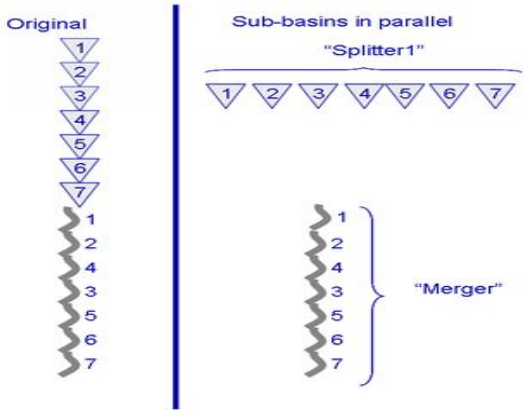


TowardOpenResourcesUsingServices

Co-funded by the Erasmus+ Programme of the European Union

VBU - Water resources management

- GRID Computing of SWAT
- SWAT Model Parallelization:
  - Split large SWAT models at sub-basin level
  - Compute them separately as independent tasks
  - Merge individual outputs from each sub-basin and route the outputs through the river network



7 sub-basins, 7 HRU's:	**Computation time (seconds)**		Number of CPUs	Speedup
Full model ("sequence")	**32**			
Parallelisation Experiment	**Approach I**	**Approach II**		
Splitting	1.2	1.4		
Sub-basin	3.3	5		
Merging	6.3	4.4		
Parallel computing	**10.8**	**10.8**	7	2.96

S. Yalew, A. van Griensven, N. Ray, L. Kokoszkiewicz, G.D. Betrie, Distributed computation of large scale SWAT models on the Grid, Environmental Modelling & Software 41 (2013) 223-230

TowardOpenResourcesUsingServices

Co-funded by the Erasmus+ Programme of the European Union

VBU - Water resources management

- **Future developments** (community/demand driven)
- STANDARDISATION for
  - Data exchange, model exchange and data-model exchange
  - Interoperability
- QUALITY CONTROL
  - Data models and metadata for observed data and model results
  - User rating
- LIBRARIES & PORTALS
  - Repositories for data, models and model applications
  - Open access

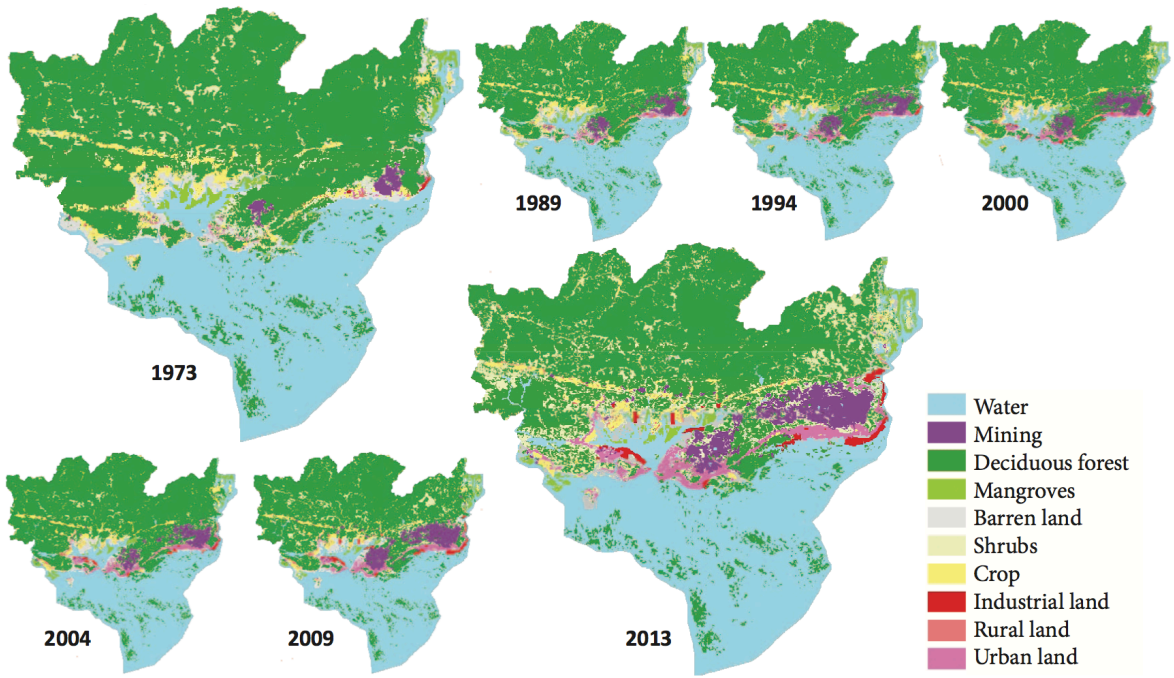
JJT2 - Statistical approach to Geography

- Dominique Laffly, Nathalie Hernandez, Florent Devin, Astrid Jourdan, Yannik Le Nir
- Toulouse University 2 and EISTI Pau

JJT2 - Statistical approach to Geography

- Understanding environment changes using statistical analysis of several datasets: satellite images, in-situ measurements, online databases, etc…
  - Land use change over time (e.g. Vietnam rural to urban areas)
  - Urban management (e.g. Predict effects of urban changes on quality of life in the city)
  - East Loven glacier mass balance in Spitsbergen - 78°N, 12°E, Svalbard, Norway (e.g. Predict evolution of glacier size/mass/etc…)
- Use of Multiple Correspondence Analysis (MCA), Agglomerative Hierarchical Clustering (AHC), Supervised Classification, …

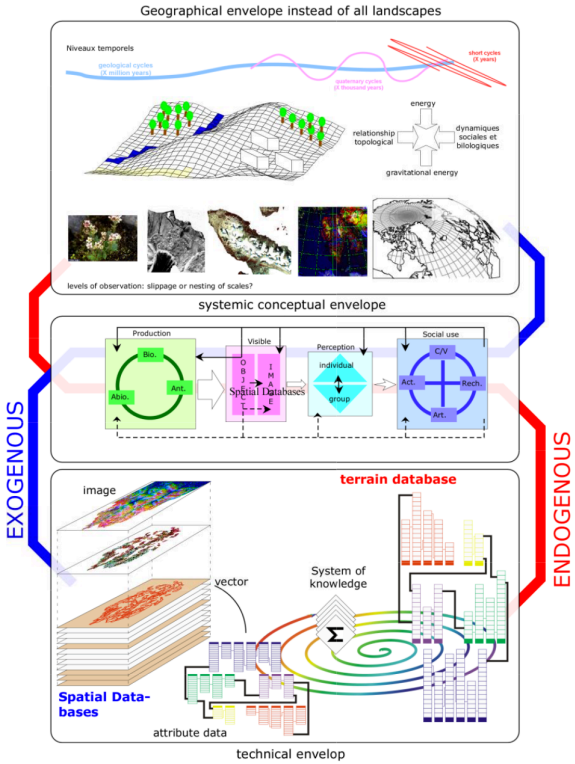
JJT2 - Statistical approach to Geography



Land use land cover patterns in the Ha Long bay area from 1973 to 2013

JJT2 - Statistical approach to Geography

Evolution of glacier correlating data coming from satellite images and in-situ monitoring



Co-funded by the Erasmus+ Programme of the European Union

JJT2 - Statistical approach to Geography

- Tools and frameworks used for data collection, storage and statistical analysis:
  - Spark
  - R
  - Scala
  - MongoDB
  - Hupi

Conclusions

- Collection of requirements is almost finalized
- Some applications are ready to be “easily” run on private or commercial clouds (standard software on SaaS/PaaS)
- For other applications (more HPC-oriented) studies are in progress
  - WRF + Chimere/CAMx
    - Optimization of use-cases workflow in private cluster and exploration of existing solutions (e.g. WRF4G and its evolution to Clouds)
  - SWAT
    - Further develop current solutions and asses performances of runs on grid
    - Evaluating existing solutions (e.g. SWAT watershed calibration on Azure, …)