ZEM: Integrated Framework for Real-Time Data and Model Analyses for Robust Environmental Management Decision Making

Velimir V. Vesselinov, Dan O’Malley, Danny Katzman

Computational Earth Science, Los Alamos National Laboratory

Waste Management Symposium, March 8, 2016
LA-UR-16-21469
ZEM framework

- ZEM provides automated and reproducible workflow interconnecting Data ⇔ Models ⇔ Decisions
- ZEM is designed for high-performance computing and big-data analysis
- ZEM employs community software (git/gitlab) for version control, team collaboration and project management using cloud-based repositories (gitlab.com / git.lanl.gov) ⇒ all past model inputs and obtained outputs are stored and can be reproduced
- ZEM provides quality assurance of the performance assessment process
- ZEM is written predominantly in julia
- julia: novel high-performance/dynamic language for technical computing (developed at MIT)
ZEM components

- **MADS** (Model Analysis & Decision Support): actively developed open-source high-performance computational framework for data- & model-based analyses in *julia* (madsjulia.lanl.gov)
- **MySQL** (www.mysql.com): open-source relational database management system stores all the site data (more than $10^7$ entries)
- Web interfaces (for data queries and exploratory model analyses)
- Various simulators
- Visualization tools (matplotlib, gnuplot, Gadfly, Paraview, VisIt)
- *julia*/Python scripts to couple all the ZEM components

For example, a single *julia* script can:
- perform automated data query from the ZEM database
- place the data in the model input files
- initiate the simulations on HPC clusters
- generate plots and movies with the final results
Analytical solutions for groundwater flow (implemented in MADS and Wells)

Analytical solutions for Fickian (classical) and non-Fickian (anomalous) contaminant transport (implemented in MADS)

Analytical simulator of groundwater flow and contaminant transport associated with infiltration recharge and perched horizons in the vadose zone (a fast screening tool) (implemented in MADS)

Semi-analytical simulator for capture zone estimation and tracer test interpretation (push-and-pull and cross-well tracer tests; MADS)

Analytical method for removal of barometric pressure and tidal effects in the water-level data (CHiPBeta):
**ZEM: Numerical simulators**

- **FEHM**: groundwater flow and contaminant transport; geochemical reactions (LANL developed code)
- **PFloTran**: groundwater flow and contaminant transport; biogeochemical reactions (LANL developed open-source code)
- **LaGriT**: grid generation (LANL developed open-source code)
- **Ashley**: particle-based geochemical reactions (LANL developed code in Julia)
- **FEniCS**: automated and efficient differential-equation solver (open-source community code)
- **libMesh**: advanced parallel partial-differential-equation solver (open-source community code)
- **Amanzi**: groundwater flow and contaminant transport; geochemical reactions (LANL developed code; future work)
ZEM: advanced data/model analysis tools

- **Drawdown estimator**: tool for data- and model-based analysis for identification and deconstruction of pumping drawdowns (typically, drawdowns are smaller than the barometric pressure fluctuations and caused by overlapping pumping events)

- **RMF (Robust Matrix Factorization)**: novel methodology for model-free inversion and data analysis

- Unsupervised objective **machine-learning methods** for data, model and decision analyses

- **Surrogate modeling** using state-of-the-art and newly developed methods (SVR, Bayesian)

- Various **data-analysis tools** such as principle and independent component analysis, trend analysis, spatial interpolation, etc. (utilizing third-party *julia* community modules).
ZEM utilizes state-of-the-art and novel advanced methods for characterization of aquifer heterogeneity

- **Pilot-point**-based methods
- **Fourier**-based stochastic methods
- **Regularization**-based methods
- **Level-set** tomography (geologic facies reconstruction)
- “**Honest**” tomography (accounting for uncertainties and unknowns)
- **Principal Component Geostatistical Analysis** (PCGA; Kitanidis et al., 2014)
- **Random Geostatistical Analysis** (RGA) for **big-data** tomography (Le et al., 2016)
ZEM: Analyses

ZEM have been successfully applied to support development of the site conceptual model representing hydrogeological and biogeochemical processes in the subsurface

- Contaminant source identification
- Contaminant source characterization (based on geochemical data and model-free inversion using unsupervised objective machine learning)
- Monitoring network design
- Evaluation of remediation scenarios
- Sensitivity and uncertainty quantification analyses
- Decision analyses
- In the last 3 years, ZEM analyses have accumulated more than 350 CPU-years of wall-clock computational time utilizing simultaneously up to 4096 processors on the LANL HPC clusters
- ... so far, all the ZEM blind predictions have been consistent with the new observations
open-source, version-controlled, high-performance computing framework implementing state-of-the-art and novel adaptive computational techniques for:

- sensitivity analysis (local / global)
- uncertainty quantification (local / global)
- optimization / calibration / parameter estimation (local / global)
- parallel Krylov-space methods for big-data analyses
- model ranking & selection
- decision analysis (GLUE, information gap, Bayesian, Bayesian - Information Gap Decision Theory (BIG-DT), Measure-Theoretic-based approaches)
- decision-based experimental design
ZEM ⇔ MADS (Model Analysis & Decision Support)

- provides **internal** coupling with analytical groundwater flow and contaminant transport solvers
- allow **external** coupling with any existing physics simulator
- coded in **julia**
- source code, examples, test problems, performance comparisons, and tutorials are available at:
  - http://madsjulia.lanl.gov
  - http://madsjl.readthedocs.org/
Probabilistic methods work very well for dice-rolling experiments.
Probabilistic methods work very well for dice-rolling experiments.
However, many earth-science uncertainties cannot be represented probabilistically (for example, using GoldSim).
Probabilistic methods work very well for dice-rolling experiments. However, many earth-science uncertainties cannot be represented probabilistically (for example, using GoldSim). Actual geologic heterogeneity is typically unknown (left die).
Probabilistic methods work very well for dice-rolling experiments.

However, many earth-science uncertainties cannot be represented probabilistically (for example, using GoldSim).

Actual geologic heterogeneity is typically unknown (left die).

We also do not know which of the possible models of geologic heterogeneity is representative (right die), but probabilistic methods require to choose a single representative model conditioned on the available data.
We also do not know what all the sides of the dice look like, and how many sides there are.
We also do not know what all the sides of the dice look like, and how many sides there are. Therefore, we cannot enumerate all possible outcomes.
We also do not know what all the sides of the dice look like, and how many sides there are.
Therefore, we cannot enumerate all possible outcomes.
All these issues make purely probabilistic analyses flawed for many earth-science problems.
We also do not know what all the sides of the dice look like, and how many sides there are. Therefore, we cannot enumerate all possible outcomes. All these issues make purely probabilistic analyses flawed for many earth-science problems. **Bayesian - Information Gap Decision Theory (BIG-DT)** for Uncertainty Quantification & Decision Analysis has been developed to address these issues (O’Malley & Vesselinov 2014 SIAM UQ).
ZEM development support

- LANL Environmental Projects
- DiaMonD Project:
  - DiaMonD: Integrated Multifaceted Approach to Mathematics at the Interfaces of Data, Models, and Decisions
    - University of Texas at Austin
    - Massachusetts Institute of Technology (MIT)
    - Stanford University
    - Colorado State University
    - Florida State University
    - Los Alamos National Laboratory
    - Oak Ridge National Laboratory
  - Funded by DOE Office of Science
  - http://dmd.mit.edu
ZEM workflow: Data ⇔ Models ⇔ Decisions

- DataBase
  - Noisy Water Levels
  - Barometric
  - Earth Tide
  - \(10^7\)datum

- Removing baro/tide effects
  - \(\sim 100\) screens

- Good Water Levels
  - \(\sim 15\) years

- Daily Supply Pumping
  - \(\sim 15\) years

- Pumping Tests
  - \(\sim 10\)

- Inverse Analysis #1
  - \(>200\) params
  - \(\sim 10^7\) targets
  - \(~10^{10}\) Runs

- Tracer Tests
  - \(\sim 5\)

- Concentration Records
  - \(\sim 7\) years

- Predictions
- Future Pumping
- Remediation
- Supply

- BIG DT
  - \(\sim 10^9\) Runs

- Parameters
  - Contaminant
  - Transport
  - Flow

- Inverse Analysis #2
  - \(>500\) params
  - \(10^5\) targets
  - \(~10^6\) Runs

HPC

Retrodictive Pumping Impacts

LANL Chromium site

Highlights
ZEM workflow: Data ⇔ Models ⇔ Decisions
ZEM workflow: Data ⇔ Models ⇔ Decisions

- Data: Noisy Level, Earth Observation, ~10^7 data points
- Models: Future, Pumping, Remediation, Supply, BIG DT, ~10^9 runs
- Decisions: Predictions, Tracer Tests, Concentration Records, ~7 years

Highlights:
- >200 parameters
- ~10^10 runs, targets
- R-44 #2
- >500 parameters
- Inverse Analysis #2, 10^5 targets
ZEM workflow: Data ↔ Models ↔ Decisions

Data
- Noisy
- Low
- squared
- Earth
- ~10^7 cells

Models
- Water
- Analysis
- ~10^10 Runs
- tracer Tests
- ~5 Concentration
- Records
- ~7 years

Decisions
- Predictions
- Future
- Pumping
- Remediation
- Supply

BIG
- DT
- ~10^9 Runs

HPC

>200 params

MADS

LANL Chromium site

ZEM ↔ MADS

Highlights
ZEM workflow: Data ⇔ Models ⇔ Decisions
Chromium site high-level summary

- High visibility project
- ~54,000 kg of Cr\(^{6+}\) released in Sandia Canyon between 1956 and 1972 (with substantial uncertainties and unknowns)
- Cr\(^{6+}\) detected above MCL (50 ppb; NM standard) at 6 monitoring wells in the regional aquifer beneath LANL
- Cr\(^{6+}\) plume size is about 2 km\(^2\) (region above MCL)
- Cr\(^{6+}\) plume is located near LANL site boundary
- Series of water-supply wells are located nearby (less than km)
- Contaminant mass distribution in the subsurface is unknown
- Contaminant source location and mass flux at the top of the regional aquifer are unknown due to complex 3D pathways through the vadose zone
- Limited remedial options due to aquifer depth (~300 m below the ground surface) and complexities in the subsurface processes
- Current conceptual model for chromium transport in the subsurface is supported by multiple lines of evidence
Chromium project goals

- **GOAL #1**: apply modeling to support *conceptualization* of the site geologic, hydrologic and biogeochemical conditions
- **GOAL #2**: perform data- and model-based *decision analyses* for chromium remediation taking into account existing processes and uncertainties/unknowns
- **Remedial scenarios**:
  - Natural attenuation (NA)
  - Enhanced attenuation (EA; biogeochemical processes)
  - Active remediation including mass removal in the vadose zone and the aquifer (*pump-and-treat*, etc.)
  - *Combinations* of all above at different times/locations
Chromium site model

- About $10^6$ computational nodes
- Representing site geology
- Including site water-supply and monitoring wells
Drawdowns from the existing supply wells
Chromium plume transients

- Model is calibrated against all the pressure and concentration transients
- ... so far, ~20 CPU-years of wall-clock computational time are accumulated
- ... additional model improvements are still needed
Chromium plume transients

LANL Chromium site

Time=2023
Chromium bio-remediation modeling (PFloTran)
Geochemical particle-based model (Ashley)

\[ A + B = C \]

- Reduction of contaminant \( B \) by injecting \( A \)
- Reduction of contaminant \( A \) by interacting with \( B \)
- \( A \) instantaneously released (500 moles)
- \( B \) uniformly distributed in the aquifer (1000 moles)
20% of A did not react
ZEM provides automated and reproducible workflow interconnecting Data ⇔ Models ⇔ Decisions using high-performance computing and big-data analysis tools.

ZEM have been successfully applied to perform various data- and model-based analyses at the LANL Chromium site.

In the last 3 years, ZEM analyses have accumulated more than 350 CPU-years of wall-clock computational time utilizing simultaneously up to 4096 processors on the LANL HPC clusters.

... so far, all the ZEM blind predictions have been consistent with the new observations.
Many uncertainties in the environmental management problems cannot be represented probabilistically.

Newly developed methodology **BIG-DT** (Bayesian-Information Gap Decision Theory) is developed to address this issue (O’Malley & Vesselinov 2014 SIAM UQ).

**BIG-DT** is applicable to any real-world engineering problems.

**BIG-DT** is available in **MADS** (open source code written in **Julia**)
- [http://madsjulia.lanl.gov](http://madsjulia.lanl.gov)
- [http://madsjl.readthedocs.org/](http://madsjl.readthedocs.org/)
Relevant Publications


Team

- Dan O’Malley
- Zhiming Lu
- Satish Karra
- Terry Miller
- Lucia Short
- Youzou Lin
- Boian Alexandrov
- Bhat Sham
- Xiaodong Zhang
- Scott Hansen
- Steve Mattis (UT-Austin)
- Matt Grasinger (Pitt)
- Ellen Le (UT-Austin)
- Justin Laughlin (UCSD)
- Natalia Siuliukina (UCSD)
- Filip Iliev (UCSC)
- Xi Chen (UT-Austin)
- Harriet Li (MIT)
- Eric Benner (UNM)
- David Barajas-Solano (UCSD)
Why ZEM?

- **ZEM ≈ ZEN**
- **ZEM**: Zeitgeist (spirit of the time) **E**nvironmental **M**odeling
- **ZEM**: the Slavic root word for Earth