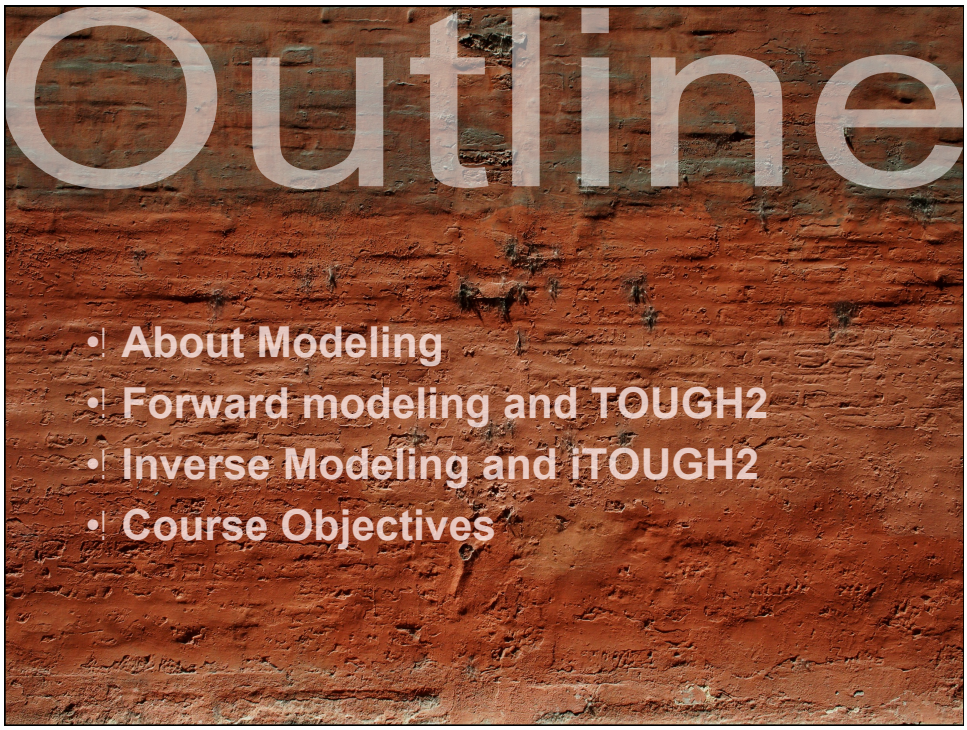
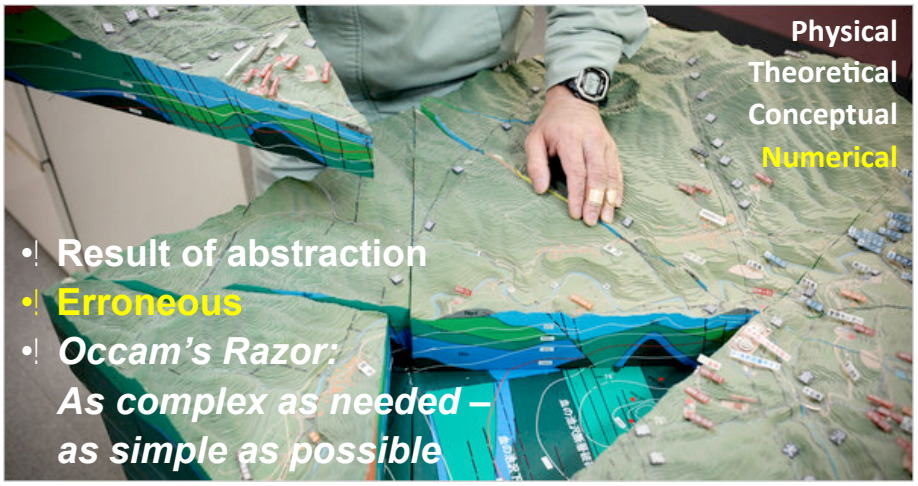
BERKELEY LAB
iTOUGH2
Short Course
Stefan Finsterle
Lawrence Berkeley National Laboratory
Earth Sciences Division
Berkeley, California


Outline
About Modeling
Forward modeling and TOUGH2
Inverse Modeling and iTOUGH2
Course Objectives

Subsurface models: then and now



"A model of the Yanaizu-Nishiyama geothermal plant. Japan's 18 geothermal plants account for only 0.3 percent of its electricity production."
A. Pollack, Japan's Nuclear Future in the Balance, *New York Times*, May 9, 2011.

A model is…

"… a *purposeful*, *simplified* representation of a real system"

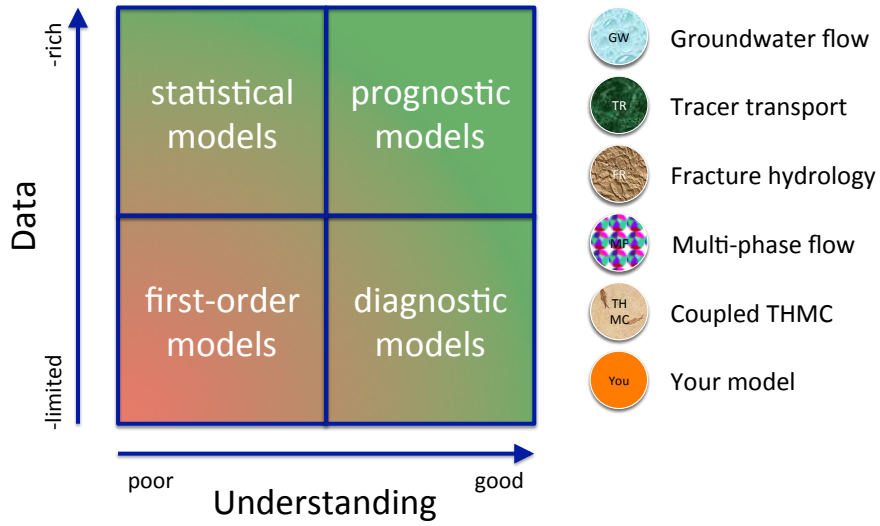
"… a simple worldview with an attitude"

"*simple and with purpose*":
The model is *simple* in that it contains only features of *primary importance* for the *intended use* of the model

Occam's razor
"Non sunt multiplicanda entia praeter necessitatem"

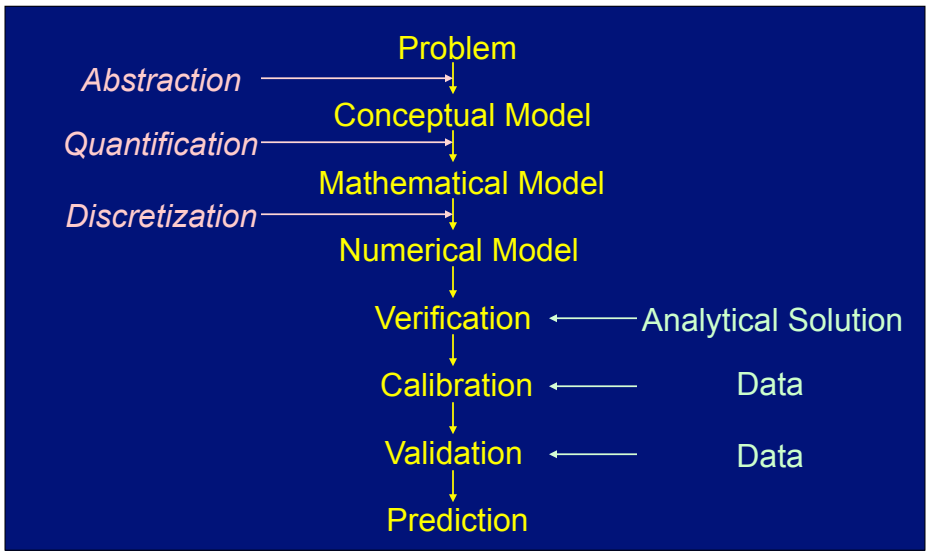
Make it as complex as needed, …
but *keep it as simple as possible!*

System understanding, data availability, and role of modeling



after Holling (1978)

Model Development



6

Role of Mathematical Modeling

Improve process understanding — TOUGH2
- Understand nonlinearities/coupled processes
- Evaluate non-observable quantities
- What-if scenarios ("virtual sandbox")

iTOUGH2

Design experiments
- Identify experimental procedure yielding data that contain information about relevant properties

Analyze data
- Determine parameters from data
- Identify model structure

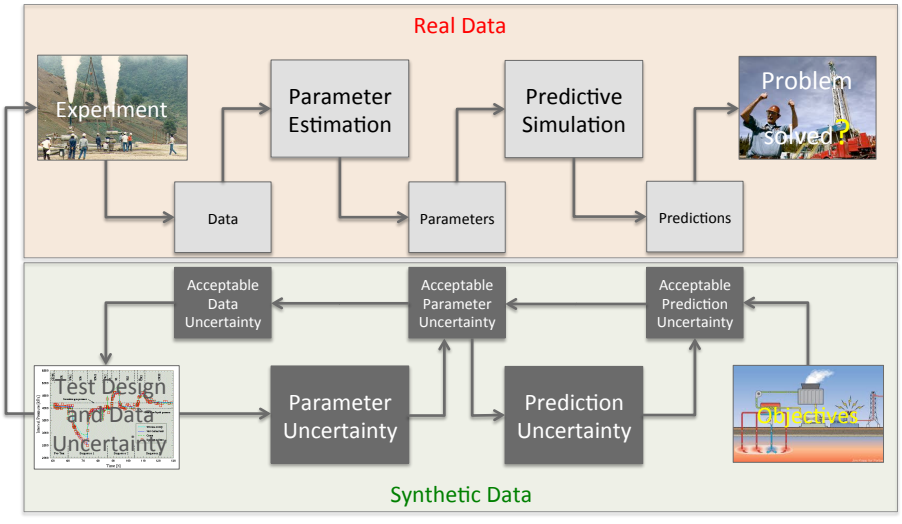
Decision support
- Risk assessment
- Sensitivity analysis

Make predictions
- Deterministic/probabilistic
- Sensitivity analysis

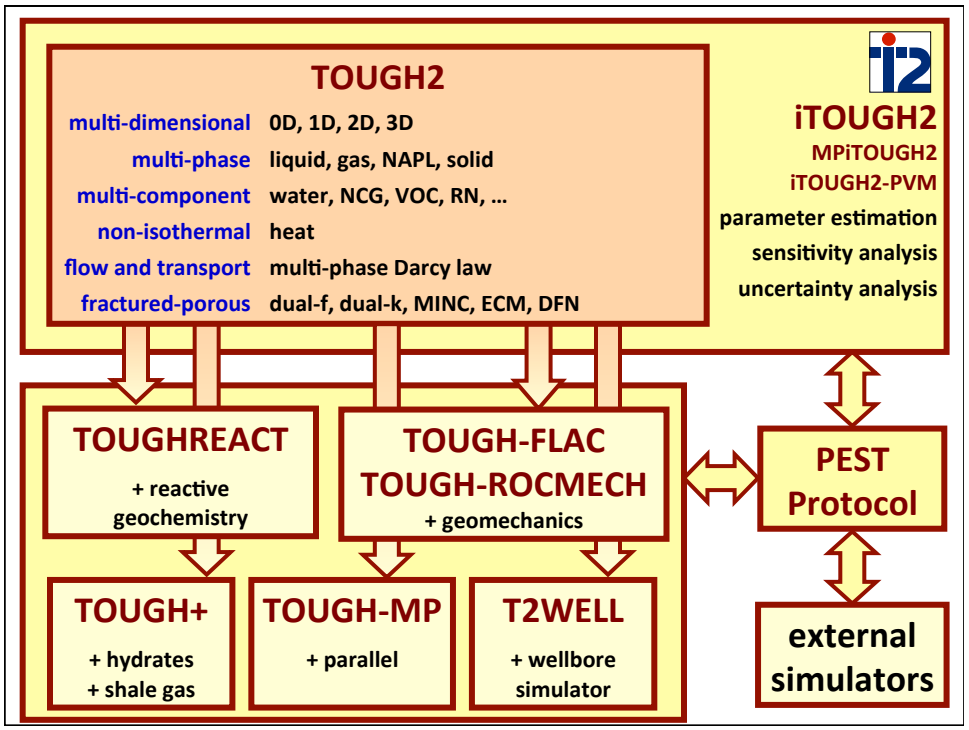
Uncertainty quantification — Optimization

7

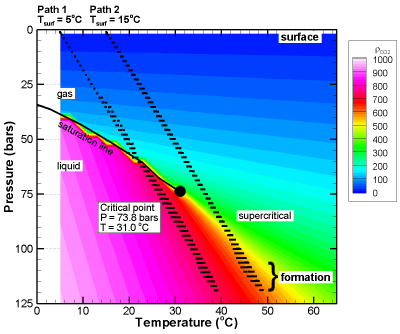
Uncertainty drives your modeling,…

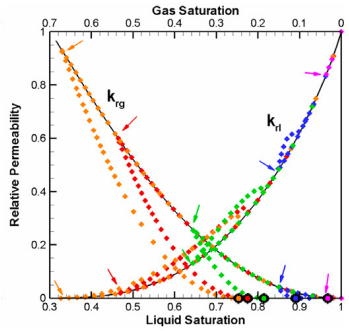


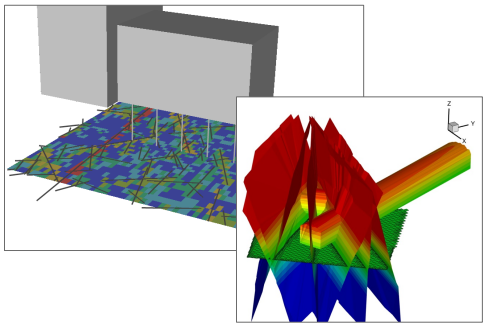
TOUGH
Suite of
Nonisothermal Multiphase Flow
and Transport Simulators


TOUGH2
multi-dimensional 0D, 1D, 2D, 3D
multi-phase liquid, gas, NAPL, solid
multi-component water, NCG, VOC, RN, …
non-isothermal heat
flow and transport multi-phase Darcy law
fractured-porous dual-f, dual-k, MINC, ECM, DFN
iTOUGH2
MPiTOUGH2
iTOUGH2-PVM
parameter estimation
sensitivity analysis
uncertainty analysis
TOUGHREACT
+ reactive geochemistry
TOUGH-FLAC
TOUGH-ROCMECH
+ geomechanics
PEST Protocol
TOUGH+
+ hydrates
+ shale gas
TOUGH-MP
+ parallel
T2WELL
+ wellbore simulator
external simulators

Components κ	Phase β			
	Solid	Fluid		
		Liquid		Gas
		Aqueous	NAPL	
Water	Ice	Liquid water	Water dissolved in NAPL	Water vapor
NCG	-	NCG dissolved in liquid water	NCG dissolved in NAPL	Non-condensible gas
VOC	-	VOC dissolved in liquid water	VOC in NAPL phase	Chem. vapor
Minerals	Grains	Dissolved, Colloids	Dissolved, Colloids	-



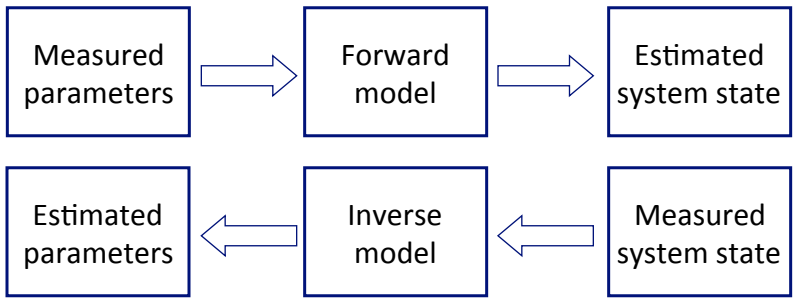




Forward Modeling

p* contains *n* measured parameters (prior information)

z contains calculated system response at *m* calibration points



r contains *m* residuals (**z*** - **z**)

p contains *n* parameters to be estimated

z* contains observations at *m* calibration points

13

- Mathematical model G relates parameters **p** to observations **z** $G(\mathbf{p}) = \mathbf{z}$
- Observations have noise:

$$\mathbf{z} = G(\mathbf{p}) + \varepsilon = \mathbf{z}_{true} + \varepsilon$$

- Forward problem: find **z** given **p**
- Inverse problem: find **p** given **z**

$$\mathbf{p} = G^{-g}(\mathbf{z})$$

- Model identification problem: find G given examples of **p** and **z**
- Discrete inverse problem: **z** is a vector of observations at discrete points in space and time ("calibration points")

14

- Inverse modeling ≈ parameter estimation ≈ model calibration ≈ history matching ≈ curve fitting ≈ regression (≈ optimization ≈ filtering)
- distinction is largely irrelevant
- preferred terms in different disciplines:
  - inverse problem: generally many parameters; ill-posed; geophysics; imaging; applied mathematics
  - parameter estimation: generally few parameters; well-posed; hydrogeology
  - model calibration: process modeling
  - history matching: reservoir engineering
  - curve fitting: data analysis
  - regression: statistics

15

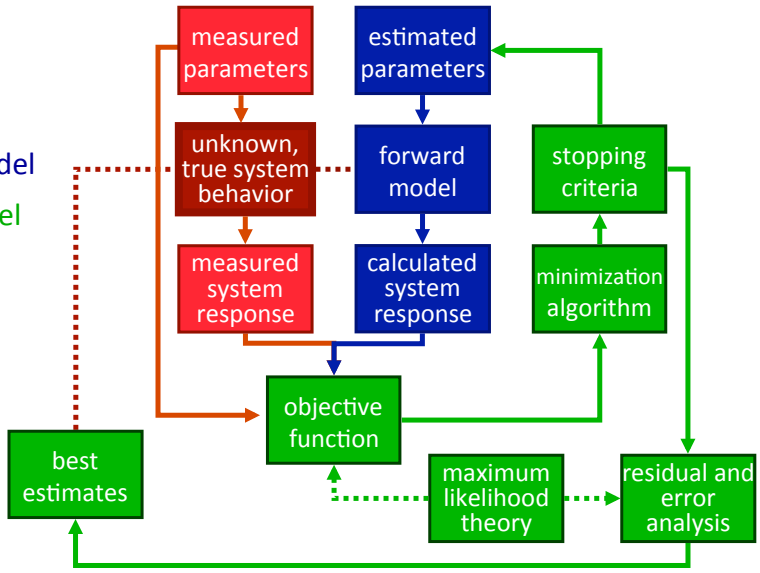
iTOUGH2 flowchart

True system

Data

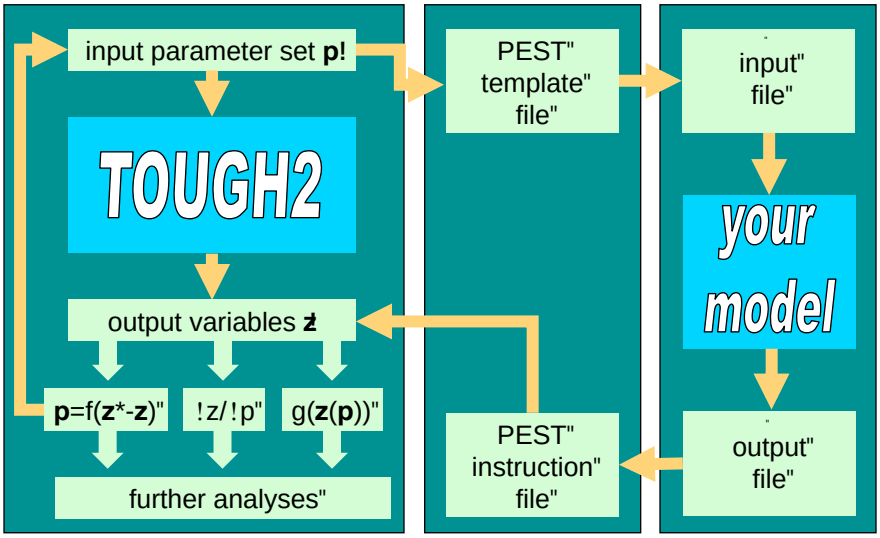
Forward model

Inverse model



16

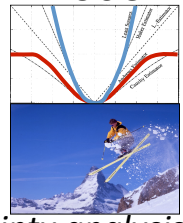
Running TOUGH2 once is tough; running it many times is iTOUGH2



17

Inverse modeling ingredients

- Parameters **p** — Parameterization
- Data **z*** — Joint inversion
- Forward operator **z**(**p**) — TOUGH2
- Objective function S(**z***-**z**(**p**))
- Minimization algorithm min S(**p**)
- Sanity check S_{min}, σ_p, σ_z *uncertainty analysis*

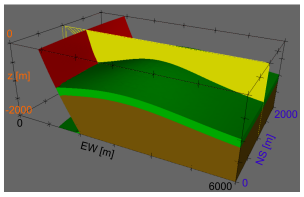


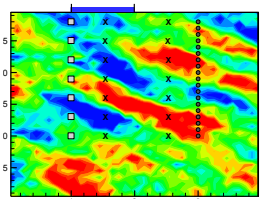
ingredient #1:

parameters

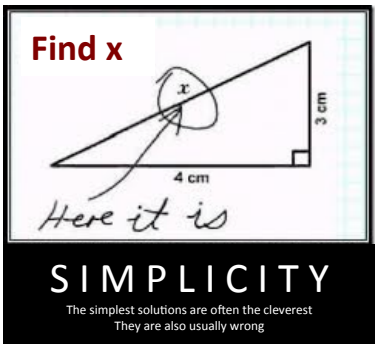
So many parameters…so little (CPU) time!

- Parameterization
  - Properties
  - Structure
  - Forcing terms
  - Conceptual errors
  - Joint estimation of properties, structure, forcing terms, errors





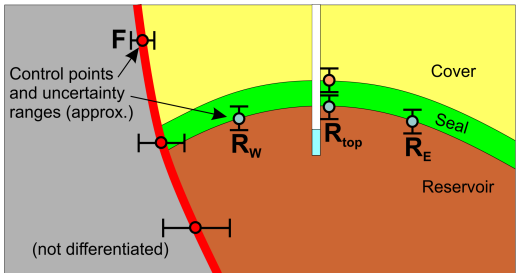
- Occam's razor
  - *as complex as needed – as simple as possible*
  - Start simple and add complexity
  - Start complex and simplify
- Parameter selection
  - Sensitivity
  - Independence
  - Superparameters

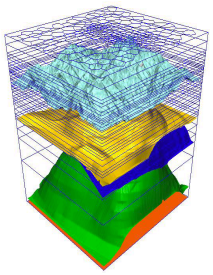


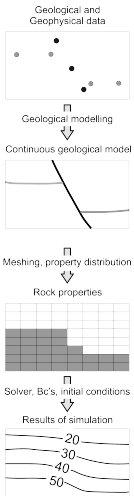
Different models created by different geologists for the same region (Courrioux et al., Experience from 3D cartography training sessions. Implication for uncertainty factors identification. 34th International Geological Congress, August 5-10, 2012, Brisbane, Australia)

Joint analysis of structural and property data reduces estimation and prediction bias

Wellmann et al., *C&G* (2014)







- Parameterize geologic structure at control points
- Develop integrated workflow from structural geologic data to discretized reservoir model
- Jointly estimate structural geometry and reservoir properties

Want more parameters?
Give me more data!

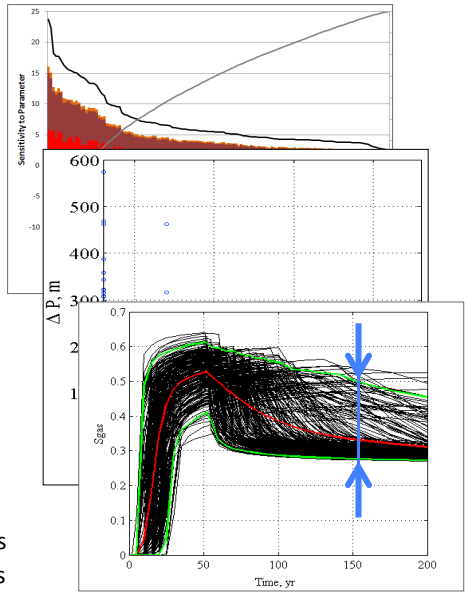
The more
high-quality,
sensitive,
complementary,
consistent
data, the better

mixing ingredients #1 and #2:

sensitivity analysis

Going global with sensitivity analysis

- Local sensitivity analysis
  - Requires (only) *n*+1 (parallel) runs
  - By-product of derivative-based minimization algorithms
  - Many useful composite measures on parameter influence and data worth
- Morris One-At-A-Time
  - Requires *r*(*n*+1) runs
  - Identifies influential and non-influential parameters
  - Identifies nonlinearity and interaction effects
- Sobol'/total sensitivity indices
  - variance/sampling based
  - Identifies contribution of uncertain parameter to prediction uncertainty
  - Fix one parameter – vary all the others
  - Vary one parameter – fix all the others

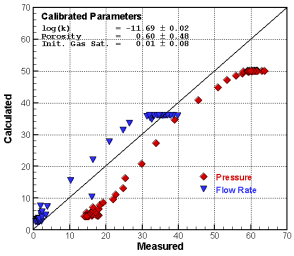


ingredient #3:

objective function

heteroscedastic

$$\tilde{z}(z;\,!) = \begin{cases} \dfrac{(z' \,!\, 1)}{!\;\; g'^{!1}} & \text{if } ! \neq 0 \\ g\;\ln(z) & \text{if } ! = 0 \end{cases}$$

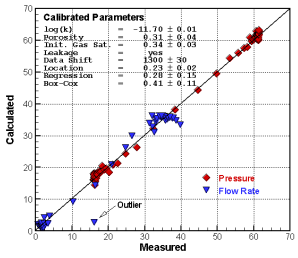


autocorrelated

$$\tilde{r}_i = \begin{cases} r_1\sqrt{(1 \,!\, !^2)} & \text{if } i = 1 \\ r_i \,!\, ! \;\; r_{i!1} & \text{if } i = 2, \ldots, m \end{cases}$$

skewed and heavy-tailed

$$p(r_i \mid !,\;\;) = \frac{2\#_{\,!}}{! + !^{-1}} \quad \exp\left\{-c_\beta \left| r_{i,t} \right|^{2/(1+\beta)}\right\}$$



concurrently estimate *statistical* parameters
β, λ, ρ, ξ

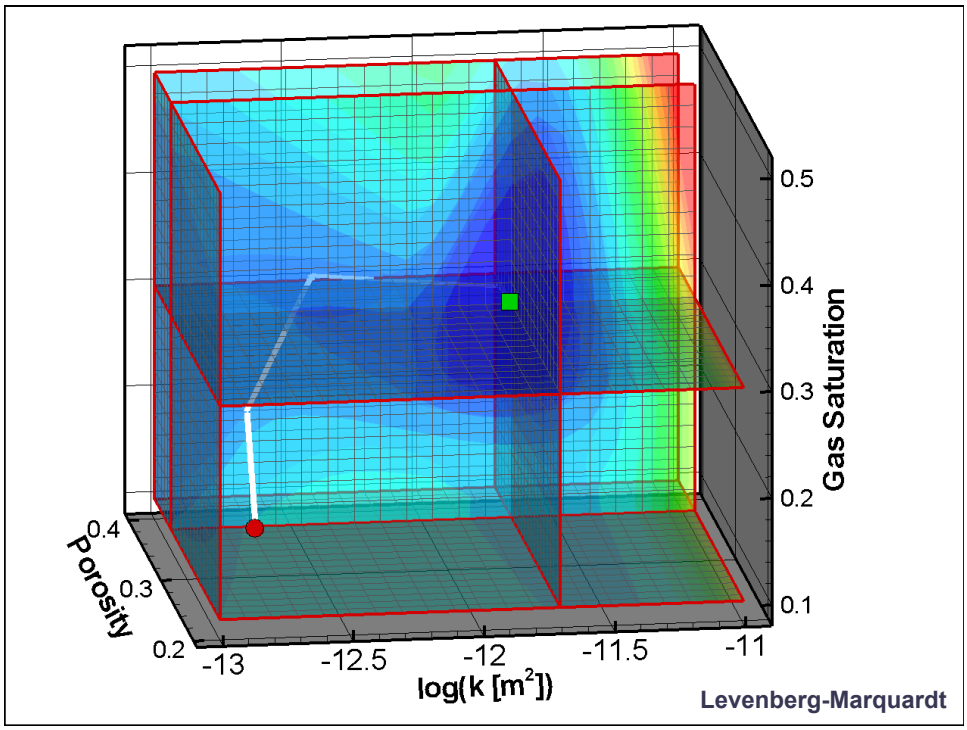
Schoups and Vrugt, *WRR*, 2010; Finsterle and Zhang, *C&G*, 2011

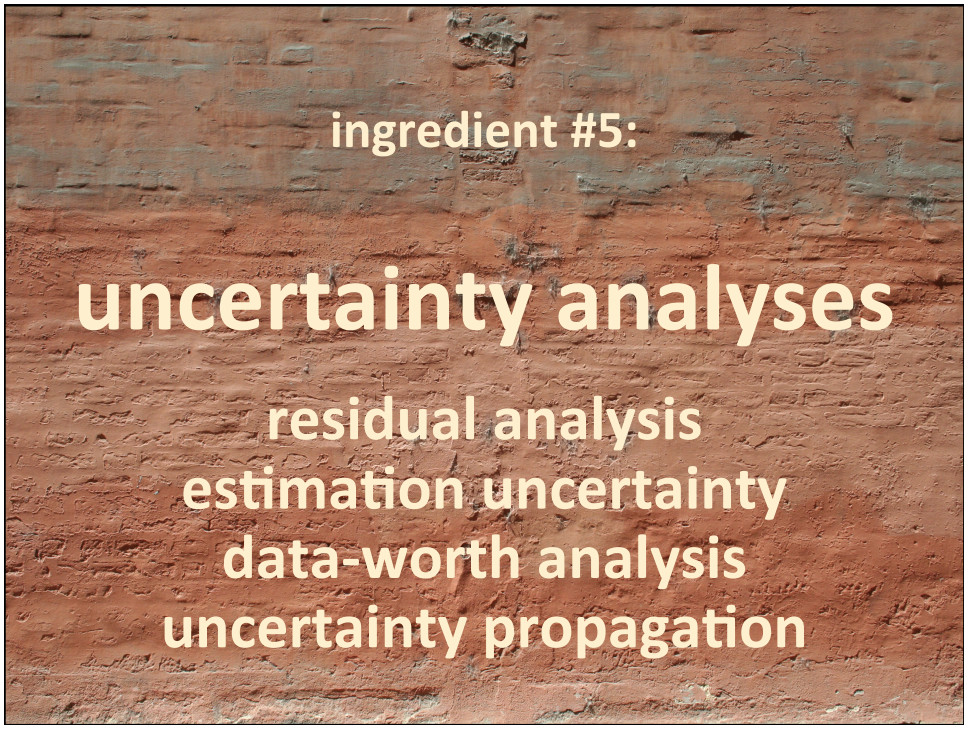
ingredient #4:

minimization algorithm

Derivative-Based Minimization Algorithms

Harmony Search
Swarm Optimization
Evolutionary Algorithms
Simulated Annealing


0.5
0.4
0.3
0.2
0.1
Gas Saturation
0.4
0.3
0.2
Porosity
-13
-12.5
-12
-11.5
-11
log(k [m²])
Levenberg-Marquardt

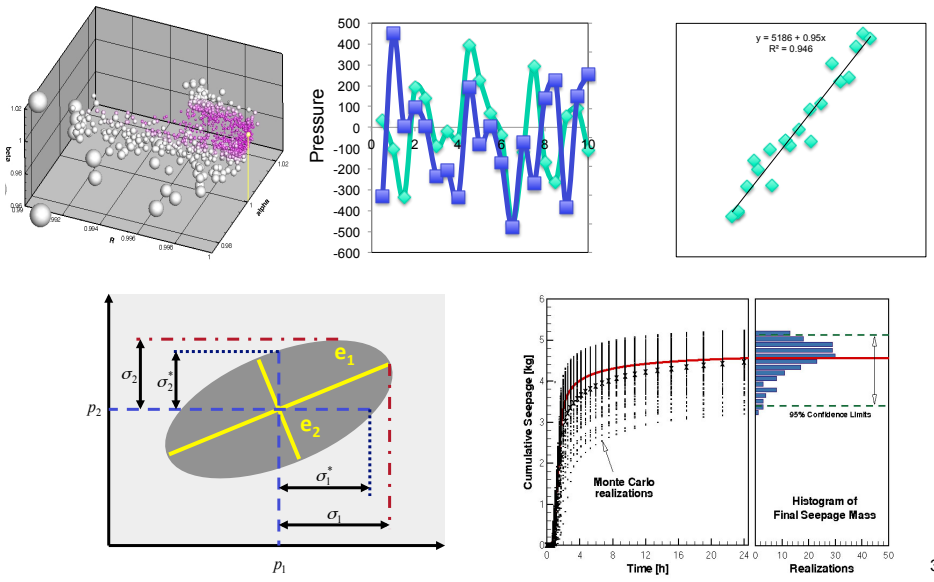


Why residual and uncertainty analysis?

- Parameter estimates may be worthless if:
  - Model does not match the data, i.e., is an unlikely representation of the true system
    goodness-of-fit, Fisher Model Test
  - Estimates are biased by systematic errors or outliers in the data
    residual analysis
  - Estimation uncertainty is large
    C_{pp}, *correlation coefficients*
  - Solution is *non-unique* or *unstable*
  - Predictions are highly *uncertain*

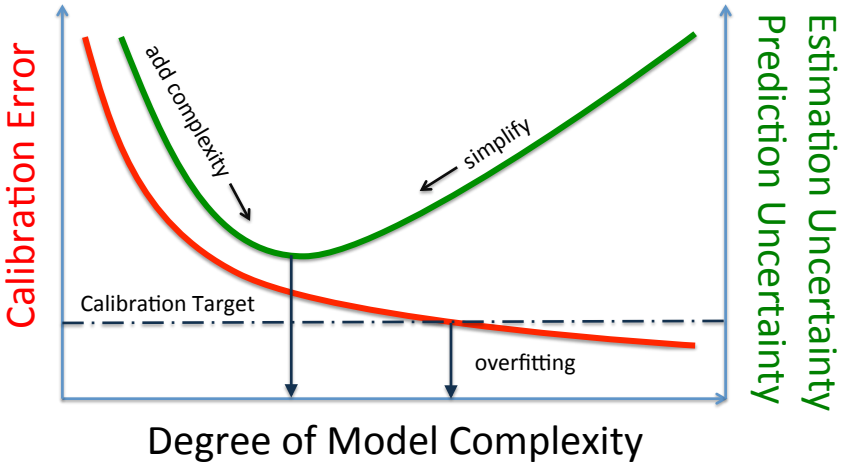
34

Residual and Uncertainty Analysis



35

Occam's Razor - sharpened!



Success Criteria

- Captures salient features of system behavior (expert judgment)
- Acceptable match (goodness-of-fit criteria)
- Acceptable estimation uncertainty (determinant of estimation covariance matrix)
- Ability to make acceptable predictions (validation acceptance criteria)
- Combination (model identification criteria)

➢ Depends on study objectives

➢ Use as criteria for test design!

37

Textbooks

- Aster et al., *Parameter Estimation and Inverse Problems*, 2nd, Ed., Academic Press, 2013.
- Hill and Tiedeman, *Effective Groundwater Model Calibration, With Analysis of Data, Sensitivities, Predictions, and Uncertainty*, Wiley, 2007.
- Saltelli et al., *Global Sensitivity Analysis, The Primer*, Wiley, 2008.
- …

38

Course Objectives

- **General:** Provide participants with *conceptual understanding*, *theoretical background*, and *practical experience* in solving simulation-*optimization* problems in geosciences and engineering using *iTOUGH2*.
- **Lectures:** Understand optimization and uncertainty quantification:
  - Fundamental concepts
  - Theoretical basis
  - Practical approaches
  - Interactive discussions
- **Computer Exercises:** Gain experience using iTOUGH2
- **Course Project:** Define and develop a simulation-optimization problem of interest to you

40

iTOUGH2 Short Course

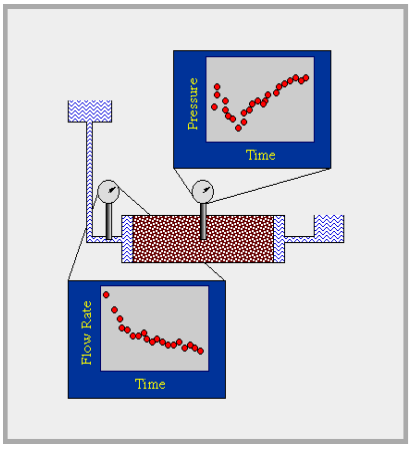
Lawrence Berkeley National Laboratory
Earth Sciences Division
Berkeley, California

Computer Exercises: Overview

Darcy
Polynomial
CO_2 Injection Test Design
Five-Spot Geothermal Injection-Production
Standard Test Set
Project

Tutorial Problem: Darcy

- Objectives:
  - Understand main iTOUGH2 concepts
  - Get familiar with key iTOUGH2 input blocks
  - Get familiar with iTOUGH2 output file
  - Examines impact of measurement noise on estimates
  - Requires some knowledge of TOUGH2 simulator
- Parameter Estimation Problem:
  - Estimate 3 parameters…
  - …based on transient pressure and flow-rate data…
  - …using Levenberg-Marquardt minimization algorithm



42

Variations of Darcy Problem

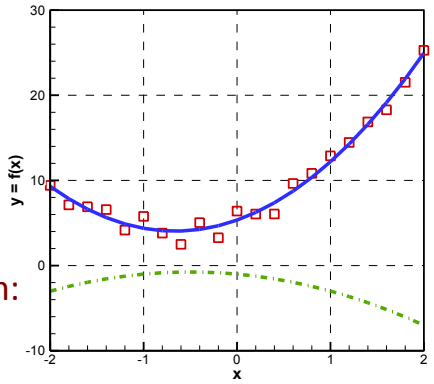
- darcy1i: run forward problem
- darcy2i: inversion with perfect data (*nonoise.dat*)
- darcy3i: inversion with noisy data (*noisy.dat*)
- darcy4i: explore
- darcy5i: grid search
- darcy6i: Morris global sensitivity analysis
- darcy7i: Saltelli/Sobol' global sensitivity analysis
- darcy8i: Monte Carlo (LHS) uncertainty quantification

43

iTOUGH2-PEST Tutorial Problem: Polynomial

- Objectives:
  - Understand main iTOUGH2 concepts
  - *Get familiar with PEST protocol*
  - Get familiar with key iTOUGH2 input blocks
  - Get familiar with iTOUGH2 output file
- Parameter Estimation Problem:
  - Estimate coefficients of polynomial…
  - …using Levenberg-Marquardt minimization algorithm

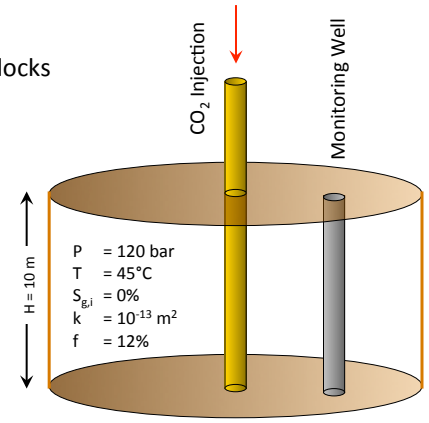
$$y(x) = \sum_{i=0}^{n} a_i \cdot x^i$$



44

Tutorial Problem: CO_2 Injection

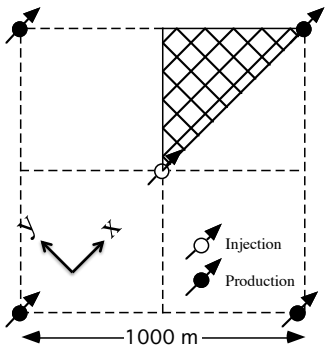
- Objectives:
  - Understand main iTOUGH2 concepts
  - Get familiar with key iTOUGH2 input blocks
  - Get familiar with iTOUGH2 output file
  - Requires some knowledge of TOUGH2 simulator (module ECO2N)
- Analyses performed:
  - Forward simulation
  - Sensitivity analysis
  - Data inversion
  - Uncertainty propagation analysis



45

Five-Spot Geothermal Injection-Production Problem

- Objectives:
  - Understand main iTOUGH2 concepts
  - Get familiar with key iTOUGH2 input blocks/ commands and output files
  - Requires some knowledge of TOUGH2 simulator (module EOS1)
- Exercises:
  1. TOUGH2 simulation with iTOUGH2
  2. Generation of synthetic data
  3. Defining parameters and performing sensitivity analysis
  4. Inversion of synthetic data
  5. Uncertainty propagation analysis
  6. Explore



46

Standard iTOUGH2 Samples and Installation Test Cases

- sample1-7: see report *iTOUGH2 Sample Problems*
- sampleLOCAL: local minimization algorithms
- sampleGLOBAL: global minimization algorithms
- sampleGS: user-specified sets (eos7c)
- sampleGSLIB: geostatistics
- sampleMOAT: Morris global sensitivity analysis
- samplePARALLEL: parallel execution
- samplePARETO: Pareto frontier using iTOUGH2-PEST
- sampleREGION: Permeability region

47

Formulate Optimization Problems

Parameters	Observations	Objective Function
Problem A:		
Problem B:		

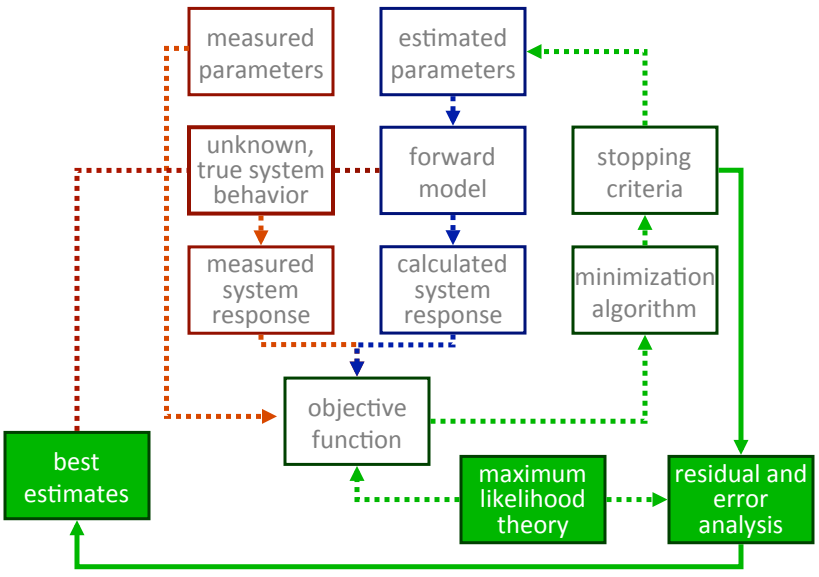
49

iTOUGH2 Short Course

Lawrence Berkeley National Laboratory
Earth Sciences Division
Berkeley, California

Residual Analysis

1

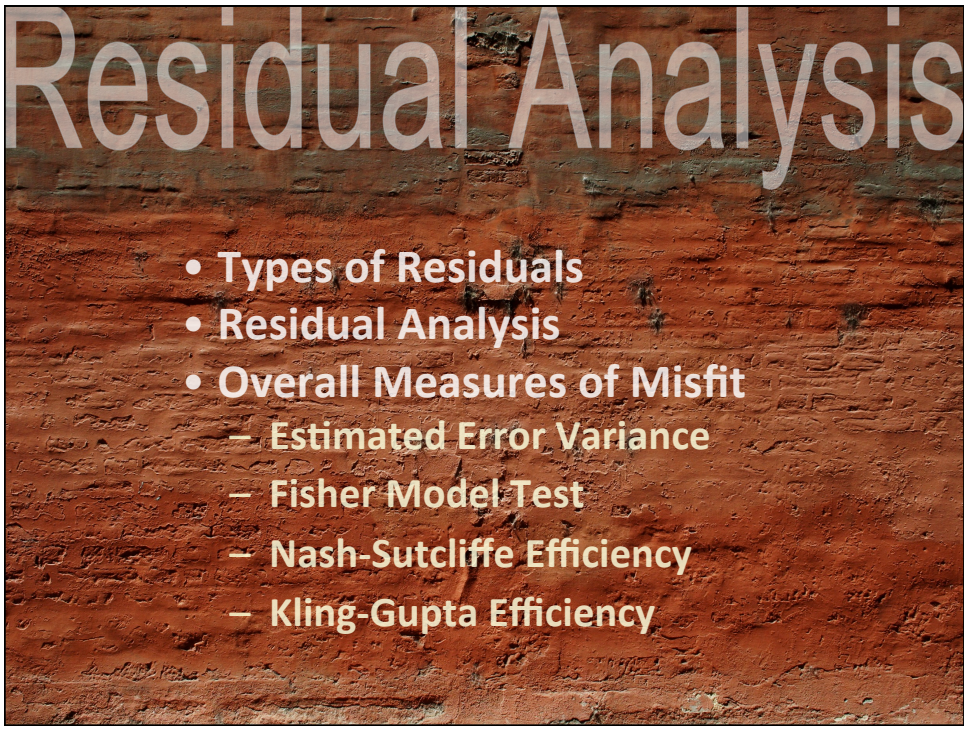


2

Why Residual and Uncertainty Analysis?

- Parameter estimates may be worthless if:
  - Model with optimized parameter does not match the data, i.e., it is an unlikely representation of the true system
    goodness-of-fit, Fisher Model Test
  - Estimates are biased by systematic errors or outliers in the data
    residual analysis
  - Estimation uncertainty is large
    C_{pp}, *correlation coefficients*
  - Solution is *non-unique* or *unstable*

3



Type of Residuals: Systematic

- Systematic error in data:
  - Data conversion error: factor, shift, drift
  - Experimental error: leak, forcing term
- Systematic error in model:
  - Inadequate process description
  - Inadequate parameterization
  - Inadequate model structure
- *Contain information* about parameters of interest
  - A function of parameters.
- *Inconsistency* between real system and its model representation
- If systematic residuals persist *after* inversion, adjust experiment *or* model!

5

Type of Residuals: Random

- Errors in stochastic model
  - Distributional assumption
  - Heteroscedasticity: non-uniform variance
  - Correlations
- Reducing errors in stochastic model
  - Analyze residuals
  - Use appropriate transformations
  - Include correlations
  - Use robust estimators

$$\tilde{z}(z;\lambda)=\begin{cases}\dfrac{(z^{\lambda}-1)}{\lambda\cdot g^{\lambda-1}} & \textbf{if } \lambda\neq 0\\ g\cdot\ln(z) & \textbf{if } \lambda=0\end{cases}$$

$$\tilde{r}_i=\begin{cases}r_1\sqrt{(1-\rho^2)} & \textbf{if } i=1\\ r_i-\rho\cdot r_{i-1} & \textbf{if } i=2,\ldots,m\end{cases}$$

$$\hat{r}=\begin{cases}1-\cos(\tilde{r}/c) & \textbf{if } |\tilde{\mathbf{r}}|\leq \mathbf{c}\pi\\ 2 & \textbf{if } |\tilde{r}|>\mathbf{c}\pi\end{cases}$$

6

Steps of Residual Analysis

① Plot the residual
→ look for randomness

② Standard statistics (mean, median, skewness, kurtosis)
→ look for symmetry

③ Regression analysis on measure vs. calculated
→ look for zero intercept, unit slope, and unit Pearson's *R* (coefficient of determination)

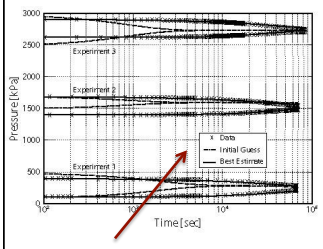
④ Fisher, Nash-Sutcliffe and Kling-Gupta efficiency criteria
→ close to 1

⑤ Runs statistics
→ Statistically tests number of positive and negative runs in residuals

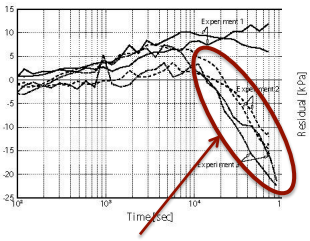
7

Remedies

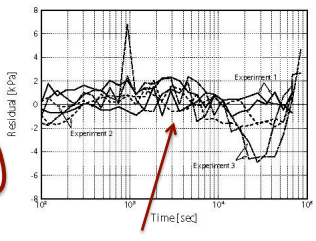
- Residuals should be *random*
- Trends and patterns in residuals indicate *systematic errors*
- Try to *remove systematic errors* by:
  - Refining the functional model
  - Correcting systematic data errors
  - Parameterizing and estimating conceptual model elements
  - Parameterizing and estimating measurement errors



Perfect match between data and model



Systematic errors from gas leakage



Random residuals after gas leakage is explicitly modeled

8

Covariance Matrices

- *A priori* error variance (assumed 1.0): σ_0^2
- *A priori* covariance matrix of observations: $\mathbf{C}_{zz} = \sigma_0^2 \cdot \mathbf{V}_{zz}$
- *A posteriori* error variance: $s_0^2 = \frac{\mathbf{r}^T \mathbf{C}_{zz}^{-1} \mathbf{r}}{m-n}$
- Covariance matrix of estimated parameters:

$$\mathbf{C}_{pp} = s_0^2 \left(\mathbf{J}^T \mathbf{C}_{zz}^{-1} \mathbf{J} \right)^{-1}$$

- Covariance matrix of calculated observation: $\mathbf{C}_{\hat{z}\hat{z}} = \mathbf{J}^T \mathbf{C}_{pp} \mathbf{J}$
- Covariance matrix of residuals: $\mathbf{C}_{rr} = \mathbf{C}_{zz} - \mathbf{C}_{\hat{z}\hat{z}}$
- *i*th diagonal element of covariance matrix: $\sigma_{z_i}^2, \sigma_{\hat{z}_i}^2, \sigma_{r_i}^2$

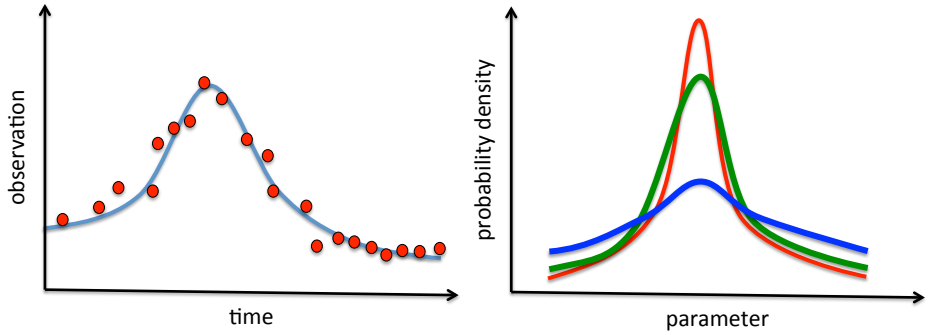
9

Residual Analysis Output in iTOUGH2

- The following are tabulated in the output file (*.out).
  - `RESIDUAL:` $r_i = z_i^* - z_i$
  - `C.O.F.:` Relative contribution [%] to objective function
  - `STD. DEV.:` *A posteriori* standard deviation of calculated system response $\mathbf{C}_{\hat{z}\hat{z}} = \mathbf{J}^T \mathbf{C}_{pp} \mathbf{J}$
  - `Yi:` Local reliability or influence ➔ $y_i = 1 - \left(\sigma_{\hat{z}_i} / \sigma_{z_i} \right)^2$
    Observations with y_i < 0.25 are poorly controlled
  - `Wi:` Studentized residual ➔ $w_i = r_i / \sigma_{r_i}$
    Potential outlier if $\text{abs}(w_i) > u(0.95) = 1.96$
  - `IOD:` Relative impact [%] of omitting observation based on D-optimality criterion (“data-worth analysis”)

10

Evaluate the worth of data points
- without having data!



$$\text{rel. data worth} = \frac{\det(\mathbf{C}_{pp}) - \det(\mathbf{c}_{pp})}{\det(\mathbf{C}_{pp})}$$

$$\mathbf{C}_{pp} = s_0^2 \left(\mathbf{J}^T \mathbf{C}_{zz}^{-1} \mathbf{J} \right)^{-1}$$

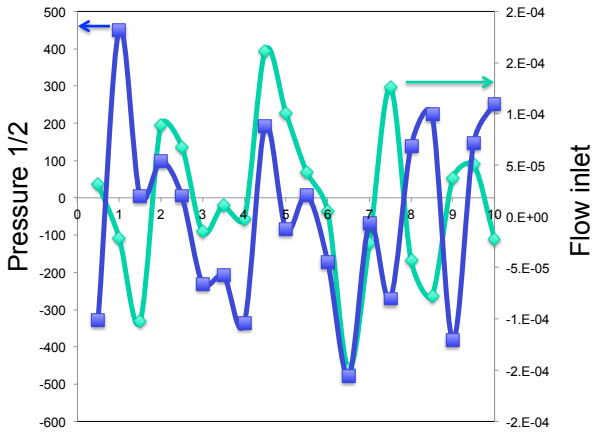
$$\mathbf{c}_{pp} = s_0^2 \left(\mathbf{j}^T \mathbf{c}_{zz}^{-1} \mathbf{j} \right)^{-1}$$

Example: darcy3i

```
*********************
 RESIDUAL ANALYSIS
*********************

[ … Description for the columns of the table … ]

------------------------------------------------------------------------
   #  OBSERVATION    AT TIME [min]     MEASURED     COMPUTED     RESIDUAL
------------------------------------------------------------------------
   1  log(abs. perm.)             -0.12000E+02 -0.11687E+02 -0.31333E+00 }
   2  Porosity                     0.25000E+00  0.37257E+00 -0.12257E+00 } prior info
   3  Initial gas sat              0.10250E+02  0.10291E+02 -0.41477E-01 }
   4  Pressure 1/2    0.50000E+00  0.10370E+06  0.10367E+06  0.35837E+02
   6  Pressure 1/2    0.10000E+01  0.10317E+06  0.10327E+06 -0.10851E+03  [cont...]
   8  Pressure 1/2    0.15000E+01  0.10257E+06  0.10291E+06 -0.33246E+03

[ … Repeated for all observations … ]
```

12

Example: darcy3i



Appear random, with no clear trend in the data

13

Summary Statistics

- `MEAN`: mean of the residual (should be close to zero)
- `STD. DEV.`: standard deviation (should be consistent with stochastic model, i.e., *a priori* defined standard deviations)
- `M/S`: Ratio of mean and standard deviation; indicates whether mean (bias) is significant (should be small)
- `SKEWNESS`: Degree of asymmetry of residuals (should be 0)
- `KURTOSIS`: Relative peakedness of distribution (should be 0)

$$SKEW = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{r_i - \bar{r}}{SDEV}\right)^3 \qquad KURT = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{r_i - \bar{r}}{SDEV}\right)^4 - 3$$

Note: statistical measures may not be robust for small data sets

14

Summary Statistics

- `MEDIAN`: median of the residuals (should be close to 0)

$$\hat{r} = \begin{cases} r_{(m+1)/2} & \text{odd } m \\ 0.5r_{m/2} + r_{m/2+1} & \text{even } m \end{cases}$$

- `AVE. DEV.`: Mean absolute deviation (should be close to 0)

$$ADEV = \frac{1}{m}\sum_{i=1}^{m}\left|r_i - \hat{r}\right|$$

- Large differences between median and mean and between standard deviation and average deviation indicate robustness issue

15

Example: darcy3i

```
Summary of Residual Analysis
----------------------------
Max weighted residual at observation    :             28
Max weighted residual                   :    -0.2300E+01
Max residual                            :    -0.4600E+03
Number of poorly controlled observations:              0
Number of large normalized residuals    :              3
Max normalized residual at observation  :             28
Max normalized residual                 :           2.40
Probable size of maximum error          :     0.5021E+03

[ ... Some iteration statistics ... ]

Control Measures
----------------
Trace (P*QLL) : n   =   3               :     0.3000E+01
Sum   (Yi)    : m-n =  37               :     0.3700E+02

Objective Function                                        C.O.F.
------------------
Initial value of objective function     :     0.5208E+03    1210.52 %
Minimum value of objective function     :     0.4303E+02     100.00 %
```

16

Example: darcy3i

[… Summary statistics for each dataset …]

```
========================================================================
DATASET              DATAPOINTS          MEAN       MEDIAN    STD. DEV.
------------------------------------------------------------------------
Pressure 1/2      [Pa]       20    -0.137E+02   -0.287E+02    0.210E+03
Flow inlet        [kg/sec]   20    -0.708E-06    0.628E-05    0.898E-04
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
ALL RESIDUALS     WEIGHTED   43    -0.359E-01   -0.000E+00    0.101E+01
========================================================================
```

Consistent with measurement noise added to data in *nonoise.dat* to create *noise.dat*

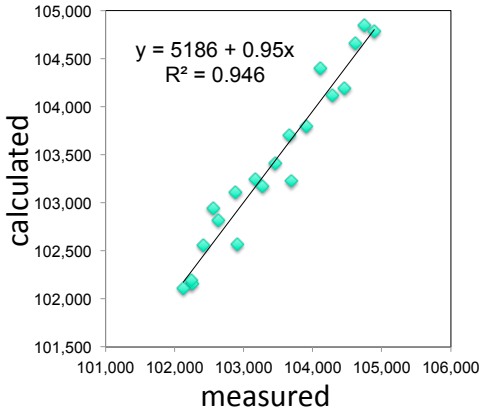
[… Summary statistics for each datatype …]

```
========================================================================
DATATYPE             DATAPOINTS          MEAN       MEDIAN    STD. DEV.
------------------------------------------------------------------------
PRESSURE          [Pa]       20    -0.137E+02   -0.287E+02    0.210E+03
FLOW RATE         [kg/sec]   20    -0.708E-06    0.628E-05    0.898E-04
========================================================================
```

17

Linear Regression Analysis

- INTERCEPT (should be 0)
- SLOPE (should be 1)
- R: coefficient of determination (should be 1)
- NSE: Nash-Sutcliffe Efficiency (should be 1)
- KGE: Kling-Gupta Efficiency (should be 1)
- GAMMAi: Contribution of correlation error, variability error and bias error to the total error



18

Nash-Sutcliffe Efficiency (*NSE*)

- A normalized model performance criterion, with observed mean as baseline

$$NSE = 1 - \frac{\sum_{i=1}^{m}(z_i - z_i^*)^2}{\sum_{i=1}^{m}(z_i^* - \mu_{obs})^2} = 1 - \frac{(\mathbf{r}^T\mathbf{r})/m}{\sigma_{obs}^2}$$

- Popular criterion in hydrology
- *NSE* ≤ 1; ideal value: *NSE* = 1
  → *NSE* can be used as objective function to be maximized
- If *NSE* ≤ 0, the model is not a better predictor than using the observed mean
- Decomposition: *NSE* = *A* – *B* – *C*

$A = R^2$: Correlation between *z*(**p**) and *z**

$B = [R - (\sigma_{calc}/\sigma_{obs})]^2$: Variability

$C = [(\mu_{calc} - \mu_{obs})/\sigma_{obs})]^2$: Bias

19

Nash-Sutcliffe Efficiency (*NSE*)

- Decomposition helps examine the different components of the residuals
- Murphy (1988) : $NSE = A - B - C$

  $A = R^2$: Correlation between *z*(**p**) and *z**

  $B = [R - (\sigma_{calc}/\sigma_{obs})]^2$: Conditional bias

  $C = [(\mu_{calc} - \mu_{obs})/\sigma_{obs})]^2$: Unconditional bias

- Gupta (2009) : $NSE = 2\alpha R - \alpha^2 - \beta^2$

  R : Correlation between *z*(**p**) and *z**

  $\alpha = (\sigma_{calc}/\sigma_{obs})$: Relative variability

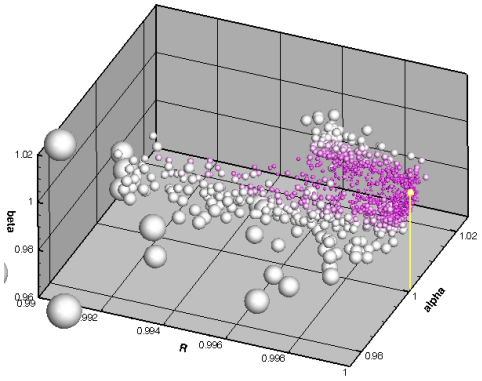
  $\beta = (\mu_{calc}/\mu_{obs})$: Relative bias

20

Kling-Gupta Efficiency (*KGE*)

- Reweights different components derived from the decomposed *NSE*
- Measures Euclidian distance from ideal point:
  $R = 1, \quad \alpha = 1, \quad \beta = 1$
- *KGE* ≤ 1
  ideal value: *KGE* = 1
  → *KGE* can be used as a multi-objective calibration criterion to be maximized

$$KGE = 1 - \sqrt{(R-1)^2 + (\alpha-1)^2 + (\beta-1)^2}$$



21

Kling-Gupta Efficiency (*KGE*)

- Relative contribution of the 3 components:

$$\Gamma_i = \frac{G_i}{G_1 + G_2 + G_3}$$

where

$$G_1 = (R-1)^2$$
$$G_2 = (\alpha-1)^2$$
$$G_3 = (\beta-1)^2$$

- For more information:
  Gupta et al., *Journal of Hydrology*, *377*, 80 - 91, 2009

22

Example: darcy3i

```
Linear Regression Analysis of Calculated Versus Observed Data ...
-------------------------------------------------------------

============================================================
DATASET              DATAPOINTS     INTERCEPT       SLOPE
------------------------------------------------------------
Pressure 1/2      [Pa]       20     0.506E+04   0.951E+00
Flow inlet        [kg/sec]   20    -0.884E-05   0.985E+00
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
ALL               WEIGHTED   40     0.366E+00   0.100E+01
============================================================


=====================================================================
       R      NSE          KGE      GAMMA1      GAMMA2      GAMMA3
---------------------------------------------------------------------
0.973       0.943       0.964       0.590       0.410       0.000
0.869       0.755       0.824       0.554       0.446       0.000
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
1.000       1.000       1.000       0.000       0.512       0.488
=====================================================+===============
```

23

Estimated Error Variance s_0^2

- The *estimated error variance*
  - a posteriori error variance
  - an aggregate measure of *goodness-of-fit*
  - represents the *mean weighted residual*

$$s_0^2 = \frac{\mathbf{r}^T \mathbf{C}_{zz}^{-1} \mathbf{r}}{m - n}$$

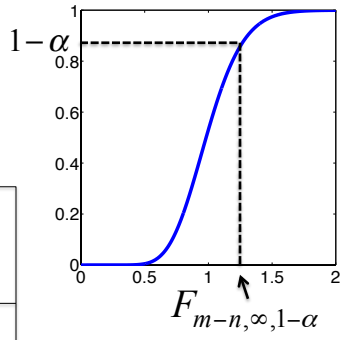
r : residual
$\mathbf{C}_{zz}$: observation covariance matrix
m : number of observations
n : number of parameters

- If *a priori* assumptions about the residuals (expressed through matrix $\mathbf{C}_{zz}$) were reasonable, then s_0^2 / σ_0^2 is close to 1: Fisher Model Test
- Recall: Solving a weighted least-squares problem minimizes the estimated error variance

24

Fisher Model Test

- The ratio s_0^2/σ_0^2 is a random variable following an *F*-distribution
- $F_{m-n,\infty,1-\alpha}$ is the inverse of the cumulative *F*-distribution for 1-α



$s_0^2/\sigma_0^2 > F_{m-n,\infty,1-\alpha}$

Error in functional or stochastic model

$1 \le s_0^2/\sigma_0^2 \le F_{m-n,\infty,1-\alpha}$

Model test passed!

$s_0^2/\sigma_0^2 \le 1$

Error in stochastic model or "overfitting"

- Only meaningful if a reliable stochastic model is available

25

Example: darcy3i vs. darcy4i

```
Fisher Model Test
-----------------
```

noisy.dat

```
Root mean square error              : 0.1078E+01
Estimated error variance            : 0.1163E+01
Variance used for error analysis    : 0.1187E+01 (a posteriori variance)
Nash-Sutcliffe efficiency criterion :     1.0000
Degree of freedom                   :         37 (no prior information)
Confidence level (1-alpha)          :       95.0 [%]
Lucky you                           : Model test successful!
                                      --> Estimated error variance is used!
Fisher model test criterion         :       1.41 (F-distribution)
Factor for confidence bands         :       2.03 (t-distribution)
Factor for confidence regions       :       2.91 (chi-square distribution)
```

noisier.dat

```
Root mean square error              : 0.3183E+01
Estimated error variance            : 0.1013E+02
Variance used for error analysis    : 0.1013E+02 (a posteriori variance)
Nash-Sutcliffe efficiency criterion :     0.9999
Degree of freedom                   :         37 (no prior information)
Confidence level (1-alpha)          :       95.0 [%]
Warning                             : Model test failed!
                                      --> Estimated error variance is used!
Fisher model test criterion         :       1.41 (F-distribution)
Factor for confidence bands         :       2.03 (t-distribution)
Factor for confidence regions       :       2.91 (chi-square distribution)
```

26

Model Identification Criteria

- Criteria to allow comparison of different inversions
  - Different data sets, parameters, or conceptual models
- Goodness of fit (s_0)
  - Not appropriate since adding more parameters always improves the fit (risk of overparameterization)
- Optimality criteria: A: trace($\mathbf{C}_{pp}$) E: max eig($\mathbf{C}_{pp}$) D: det($\mathbf{C}_{pp}$)
- Akaike Information Criterion

$$AIC = (m-n)s_0^2 + \ln|\mathbf{C}_{zz}| + m\ln(2\pi) + 2n$$

- Akaike's Bayesian Information Criterion

$$AIC = (m-n)s_0^2 + \ln|\mathbf{C}_{zz}| + m\ln(2\pi) + \ln|\mathbf{J}^T\mathbf{C}_{zz}^{-1}\mathbf{J}|$$

- Kashyap Criterion

$$d_k^* = (m-n)s_0^2 + \ln|\mathbf{C}_{zz}| + m\ln(2\pi) + n\ln(m/2\pi) + \ln|\mathbf{J}^T\mathbf{C}_{zz}^{-1}\mathbf{J}|$$

27

Example: darcy3/4i

$$c_{ij}^* = \frac{c_{ij}}{p_i p_j}$$

```
Optimality Criteria  [noisy.dat]              unscaled          scaled
-------------------
D-optimality = det(Cpp)              :      0.2622E-12      0.1306E-15
A-optimality = trace(Cpp)            :      0.5905E-03      0.2072E-02
E-optimality = max eigenvalue        :      0.5323E-03      0.2129E+00
Log-likelihood ln(L)                 :      0.2362E+02
Akaike  =-2ln(L)+2n                  :     -0.4123E+02
ABIC    =-2ln(L)+ln|F|               :     -0.2168E+02
Kashyap =-2ln(L)+ln|F|+n*ln(m/2Pi)   :     -0.1591E+02
```

negative log-likelihood to be minimized

```
Optimality Criteria [noisier.dat]            unscaled          scaled
-------------------
D-optimality = det(Cpp)              :      0.1493E-09      0.1139E-12
A-optimality = trace(Cpp)            :      0.4288E-02      0.1812E-01
E-optimality = max eigenvalue        :      0.3789E-02      0.1601E+00
Log-likelihood ln(L)                 :     -0.1423E+03
Akaike  =-2ln(L)+2n                  :      0.2906E+03
ABIC    =-2ln(L)+ln|F|               :      0.3081E+03
Kashyap =-2ln(L)+ln|F|+n*ln(m/2Pi)   :      0.3139E+03
```

All criteria have lower values for noisy.dat because of higher data quality, making it the preferred model

28

Questions Residual Analysis

1. The *a posteriori* error variance s_0^2 turns out to be significantly greater than the *a priori* error variance σ_0^2.
   - What does "significantly" mean?
   - What does that result indicate?
2. What are you looking for when evaluating residuals?

29

iTOUGH2 Short Course

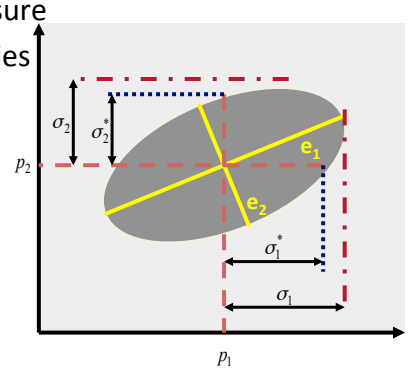
Lawrence Berkeley National Laboratory
Earth Sciences Division
Berkeley, California

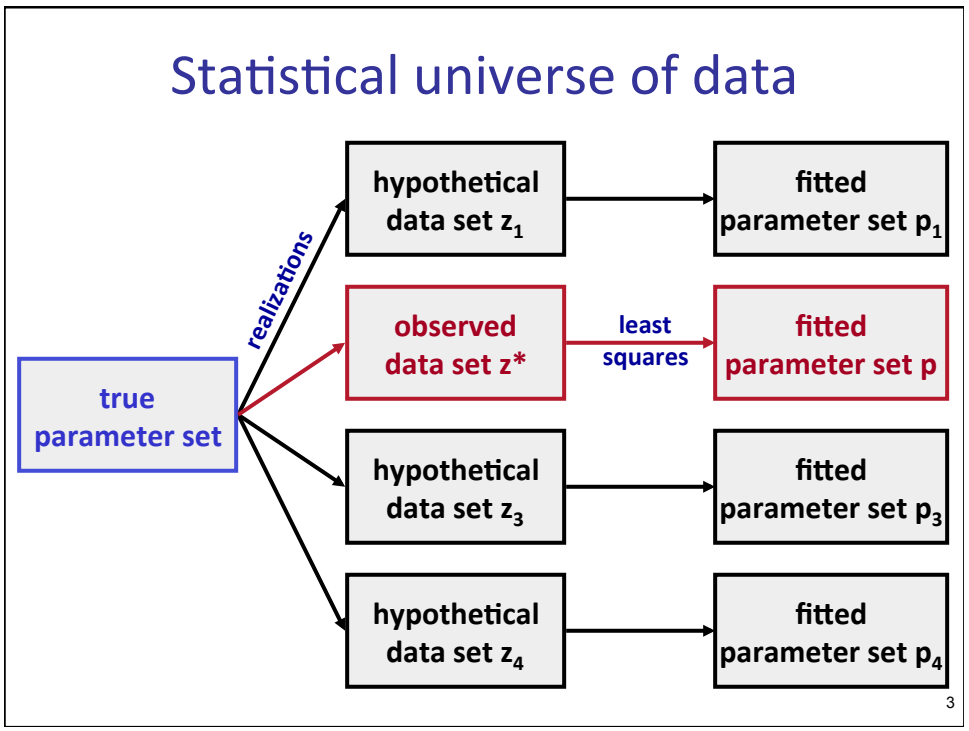
Uncertainty Analysis

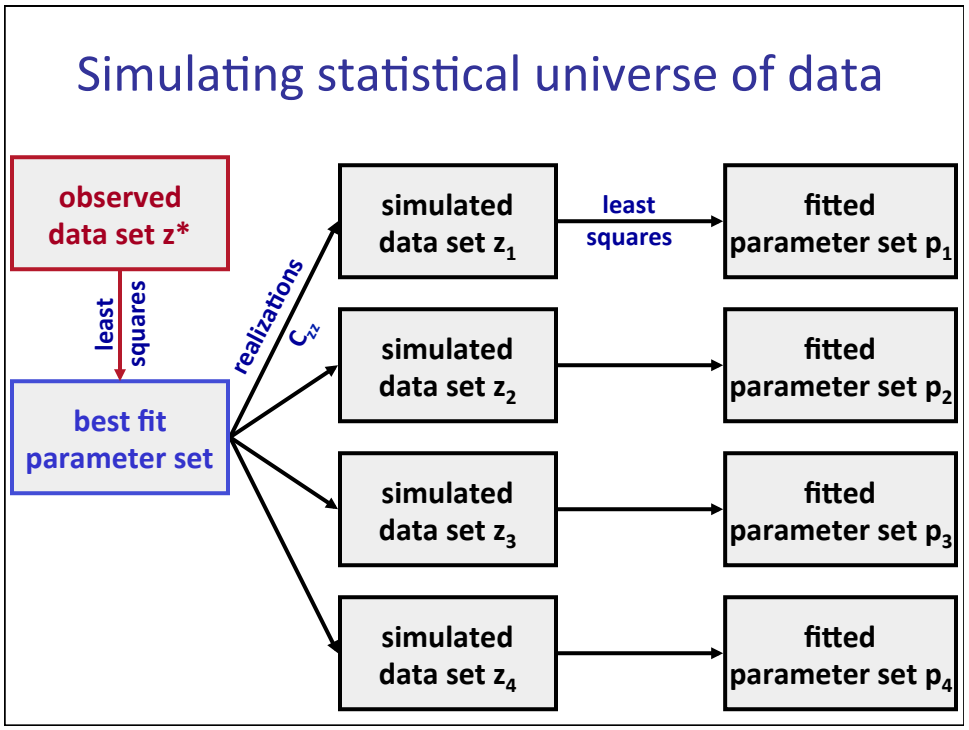
Covariance Matrix of Estimated Parameters C_{pp}
Uncertainty Propagation Analysis

Parameter uncertainty analysis in

- Covariance/correlation matrix
- Correlation chart
- Direct (pairwise) parameter correlations
- Conditional estimation uncertainties
- Overall parameter correlation measure
- Improvement over prior uncertainties
- Eigenanalysis of covariance matrix
- Correction for nonlinearities

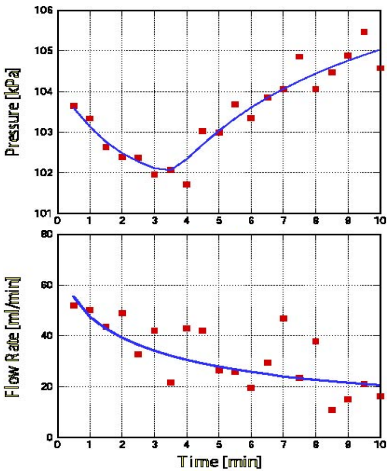



Statistical universe of data
hypothetical data set z_1
fitted parameter set p_1
realizations
observed data set z*
least squares
fitted parameter set p
true parameter set
hypothetical data set z_3
fitted parameter set p_3
hypothetical data set z_4
fitted parameter set p_4
3

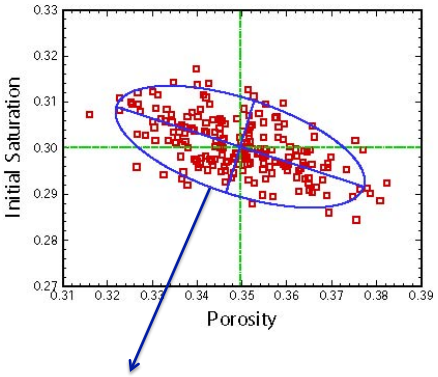

Simulating statistical universe of data
observed data set z*
least squares
best fit parameter set
realizations
C_{zz}
simulated data set z_1
least squares
fitted parameter set p_1
simulated data set z_2
fitted parameter set p_2
simulated data set z_3
fitted parameter set p_3
simulated data set z_4
fitted parameter set p_4

Simulating estimation uncertainty

200 realizations of hypothetical pressure and flow rate data



200 best estimates of porosity and initial gas saturation



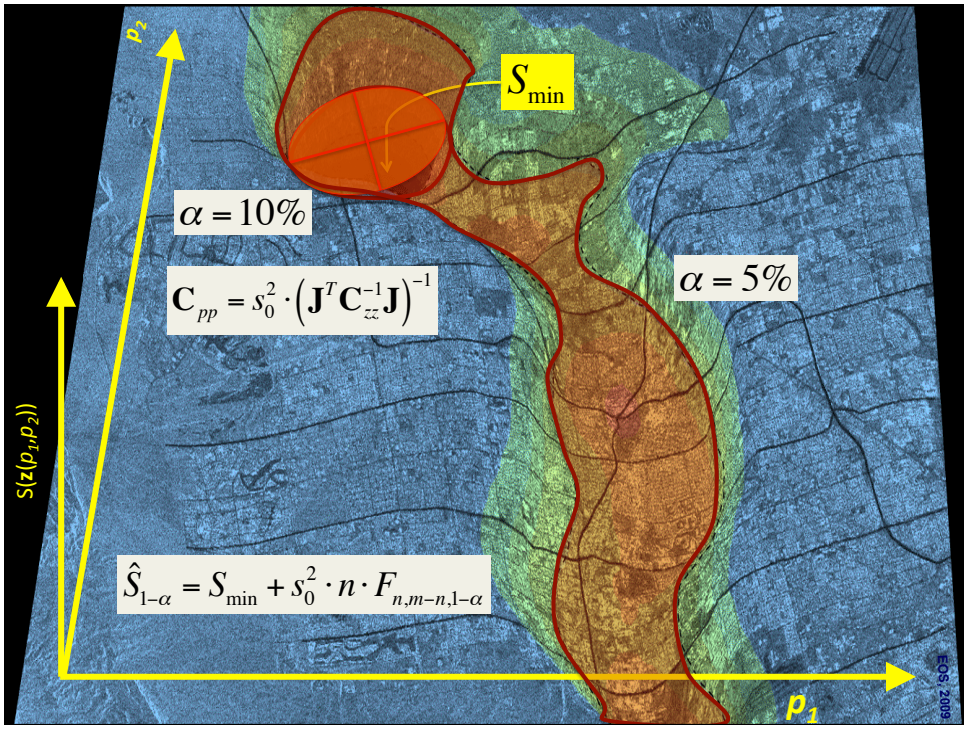
$\mathbf{C}_{pp}$ gives reasonable estimate of uncertainty region

5

Confidence Region

- The contours of the objective function visualize:
  - Confidence region
  - Correlation structure
  - Appropriateness of linearity assumption
  - Appropriateness of normality assumption
  - Well-posedness of inverse problem

6



Confidence Region

- The probability that the true parameter set lies within the ellipsoidal confidence region represented by $\mathbf{C}_{pp}$ is (1-α).
- The true confidence region is bounded by the contour line of the objective function at level:

$$S_{\min} + s_0^2 \cdot n \cdot F_{m-n,\infty,1-\alpha}$$

- The confidence region increases with decreasing α.
- $\mathbf{C}_{pp}$ is only a good representation of actual confidence region if linearity and normality assumptions are not violated.

8

Covariance Matrix of Estimated Parameters $\mathbf{C}_{pp}$

- The covariance matrix $\mathbf{C}_{pp}$ is an estimate of the uncertainty of the estimated parameters:

$$\mathbf{C}_{pp} = s_0^2 \left(\mathbf{J}^T \mathbf{C}_{zz}^{-1} \mathbf{J} \right)^{-1}$$

- $\mathbf{C}_{pp}$ is an approximation of the actual parameter uncertainty; it is based on a normality and linearity assumption
- $\mathbf{C}_{pp}$ is proportional to goodness-of-fit (s_0^2)
- $\mathbf{C}_{pp}$ is inversely proportional to sensitivity matrix ($\mathbf{J}$)
- $\mathbf{C}_{pp}$ is proportional to measurement uncertainty ($\mathbf{C}_{zz}$)

9

Estimation Uncertainty

- Decreases with improvement of fit
  - *Use good data and good model*
- Decreases with increasing sensitivity
  - *Use sensitive data*
- Decreases with decreasing correlations
  - *Use data that allow for independent determination of each parameter*
  - *Avoid overparameterization*
- Design tests accordingly!

10

Covariance Matrix
Correlation Matrix

- The diagonal of $\mathbf{C}_{pp}$ contains variances σ_{ii}^2 of estimated parameters.
- Off-diagonal elements are covariances c_{ij} between pairs of parameters.
- Off-diagonal elements are "normalized" to yield *correlation coefficients* r_{ij}:

$$-1 < r_{ij} = \frac{c_{ij}}{\sqrt{\sigma_{ii}^2 \cdot \sigma_{jj}^2}} < 1$$

11

Correlation Coefficients

- A correlation coefficient of *zero* indicates that the two parameters can be estimated *independently*.
- A correlation coefficient of *–1 or 1* indicates *non-uniqueness*.
- A negative correlation coefficient indicates that a statistically similar match can be obtained by increasing one parameter and decreasing the other.
- If correlations exist, the uncertainty in one parameter affects the uncertainty in the other parameters.
- Design experiment as to minimize correlations.

12

Example: darcy3i

```
****************
 ERROR ANALYSIS
****************

Error analysis is based on >>> a posteriori <<< variance:  1.1628577E+00

Covariance(L+D)/Correlation(U) Matrix of Estimated Parameters
------------------------------------------------------------
                  log(abs. perm.)        Porosity  Initial gas sat

log(abs. perm.)       2.77543E-04           0.843           -0.259
Porosity              2.35163E-04      2.80104E-04           -0.490
Initial gas sat      -1.74370E-05     -3.32081E-05      1.63676E-05


Matrix of Direct Correlations
-----------------------------
                  log(abs. perm.)        Porosity  Initial gas sat
log(abs. perm.)             1.000           0.851           0.331
Porosity                    0.851           1.000          -0.525
Initial gas sat             0.331          -0.525           1.000
```

Correlation coefficient, r_{ij}

σ_p^2

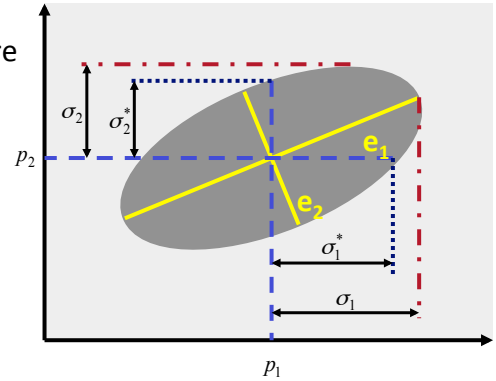
covariance, c_{ij}

Indicates degree to which a change in one parameter can be compensated by another

13

Overall Correlation

- The *conditional* standard deviation σ^* is the estimation uncertainty assuming that all other parameters are perfectly known.
- C_{pp} holds the *marginal* standard deviations σ.
- The ratio σ^*/σ is a measure of overall correlation.
- The ratio σ^*/σ should be close to 1.



14

Example: darcy3i

Arrange in order of decreasing independence (C/M)

Square root of the diagonal of C_{pp}

```
Standard Deviations
-------------------
[ ... description of the column ... ]

PARAMETER        BEST ESTIMATE A PRIORI CONDITIONAL    MARGINAL   C/M 1-M/P
Initial gas sat    1.02915E+01    N/A   3.32706E-03 4.04569E-03 0.822 1.000
log(abs. perm.)   -1.16866E+01    N/A   8.44619E-03 1.66596E-02 0.507 1.000
Porosity           3.72609E-01    N/A   7.65512E-03 1.67363E-02 0.457 1.000
```

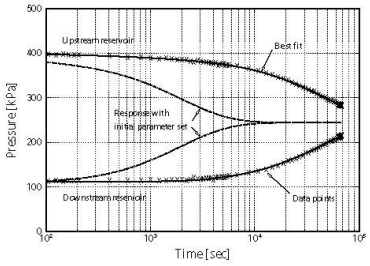
Specified only if used to weight parameters in objective function

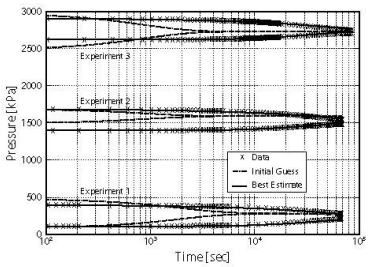
Measure of how much is learned from inversion compared to a priori uncertainty

```
Correlation Chart
-----------------

-0.6    -0.4    -0.2   0.0                0.0     0.2     0.4     0.6     0.8
|-------|-------|-------|    PARAMETER      |-------|-------|-------|-------|
     |--------|---------|  Initial gas sat
     |        |
     |        |---------|  log(abs. perm.)  |--------------------------------|
     |                                                                       |
     |------------------|     Porosity      |--------------------------------|
```

Correlation between gas saturation and porosity

Reducing Correlations: Example





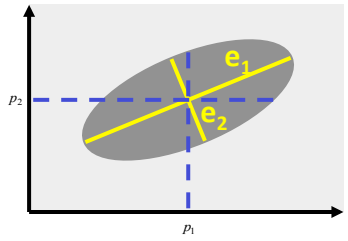
Add experiments on different pressure levels and invert jointly

	log(*k*)	log(*b*)	ϕ
log(*k*)	1.67	-0.99	-0.87
log(*b*)	-0.99	2.16	0.87
ϕ	-0.87	0.87	0.003

	log(*k*)	log(*b*)	ϕ
log(*k*)	1E-4	-0.52	-0.12
log(*b*)	-0.52	4e-4	-0.02
Φ	-0.12	-0.02	0.01

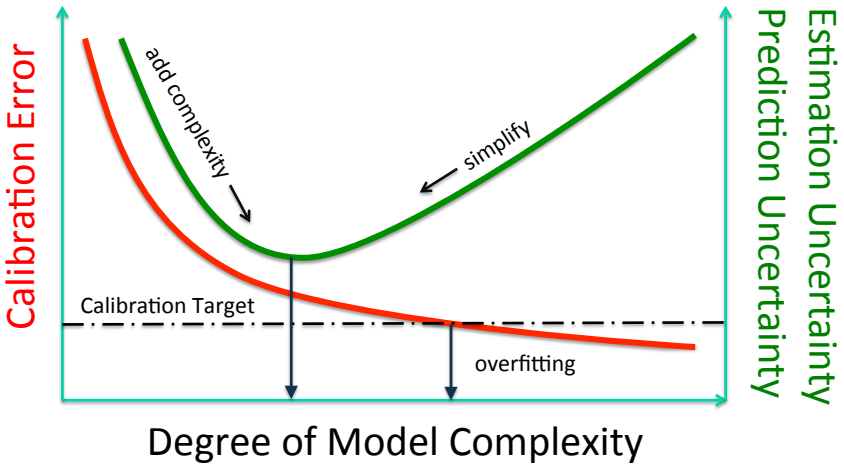
Example: darcy3i

```
Eigenanalysis of Covariance Matrix (R-Mode Factor Analysis)
------------------------------------------------------------
Performance index       :  3.25365122E-02
Condition number        :  5.07978268E+01

Eigenvalues
-----------
                                  1               2               3
  1 Eigenvalue     :        4.7286641E-05   5.1655936E-04   1.0168926E-05

Eigenvectors
------------
                                  1               2               3
  1 log(abs. perm.)        6.9034965E-01   7.0218708E-01  -1.7421439E-01
  2 Porosity              -6.5282457E-01   7.0839235E-01   2.6832882E-01
  3 Initial gas sat        3.1182918E-01  -7.1509271E-02   9.4744340E-01
```



17

Under- and overparameterization: Occam's Razor



18

Overparameterization

- A match can *always* be improved by adding more parameters to **p**.
- Adding new parameters increases correlations and thus *increases estimation uncertainty*.
- Check $\mathbf{C}_{pp}$ for large variances, correlation coefficients close to –1 or 1, and large condition numbers.
- Add parameters only if the fit can be *significantly improved* without introducing strong parameter correlations.
- *Avoid over- **and** under-parameterization*

19

Uncertainty Analysis: Questions

1. Discuss

$$\mathbf{C}_{pp} = s_0^2\left(\mathbf{J}^T\mathbf{C}_{zz}^{-1}\mathbf{J}\right)^{-1}$$

2. Under which conditions is $\mathbf{C}_{pp}$ a good approximation of the actual confidence region?
3. How can you reduce estimation uncertainty?

20

Uncertainty Analysis: Questions

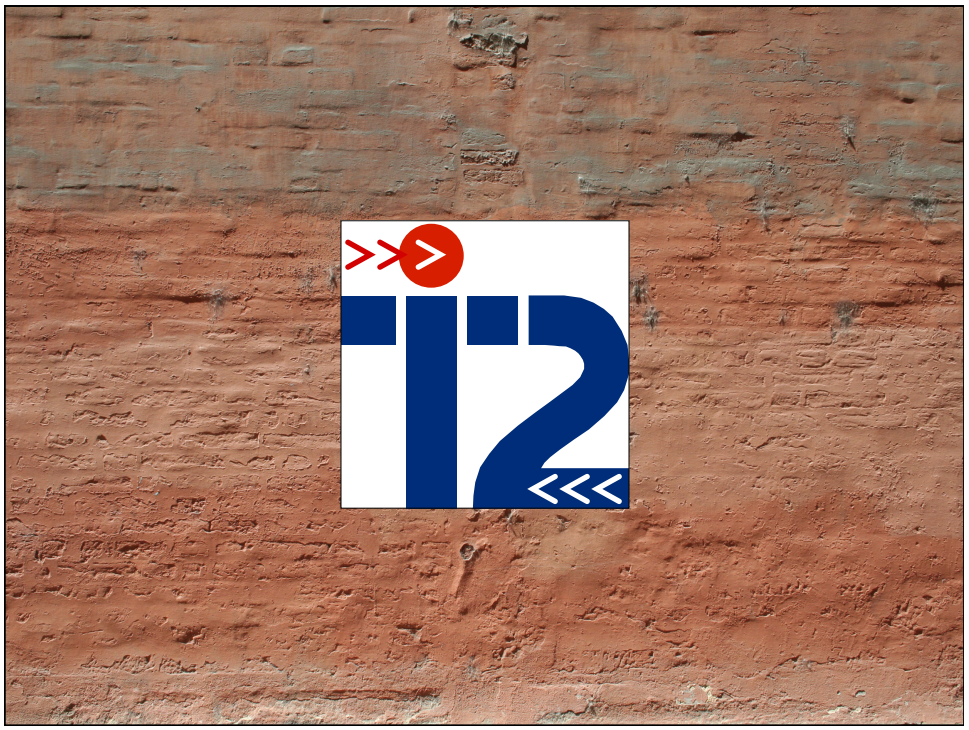
4. Discuss "underparameterization" and "overparameterization".

5. $\mathbf{C}_{zz} = \sigma_0^2 \mathbf{V}_{zz}$ is the *a priori* observation covariance matrix.

   If all elements of $\mathbf{C}_{zz}$ were multiplied by a factor of 4, how would this affect:

   - The value of the objective function S?
   - The estimated parameter set $\mathbf{p}$?
   - The estimated error variance s_0^2?
   - The uncertainty of the estimated parameters $\mathbf{C}_{pp}$?
   - The outcome of the Fisher Model Test?

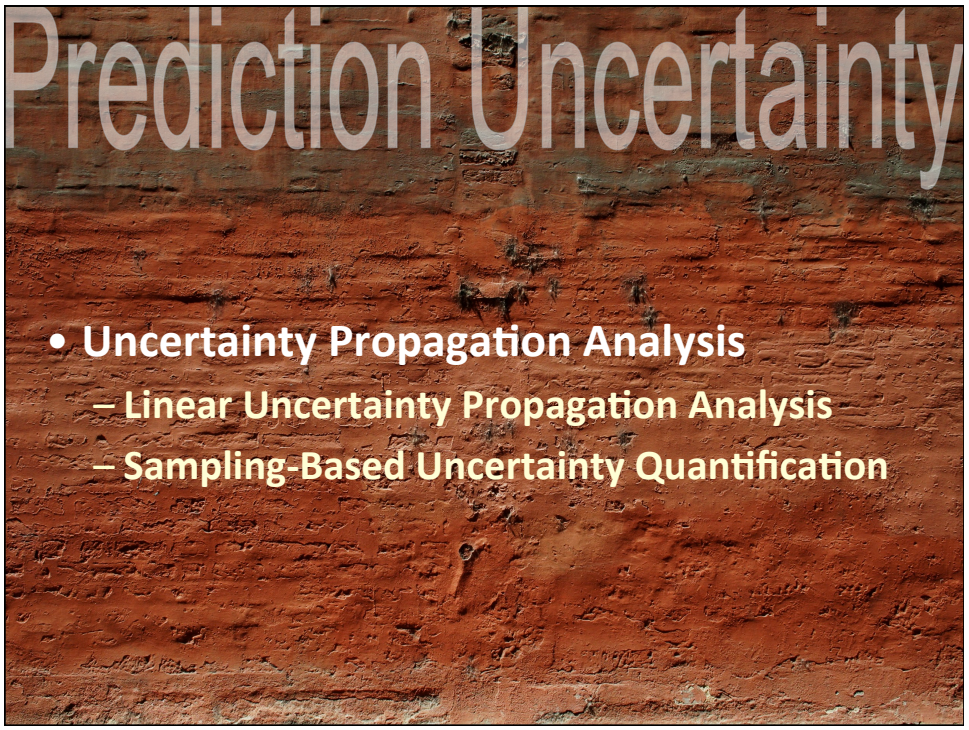
21



iTOUGH2 Commands

```
> COMPUTATION
  >> ERROR
     >>> ALPHA: alpha          significance level
     >>> A PRIORI              use σ0²
     >>> A POSTERIORI          use s0²
     >>> let FISHER model test decide
     >>> check LINEARITY assumption
```

23



Uncertainty Propagation Analysis

- Calculate prediction uncertainty as a result of parameter uncertainty.
- *Linear analysis* (First-Order Second-Moment)
  - Fast (*n*+1 forward runs)
  - Easy to report (mean and covariance matrix)
  - Based on linearity and normality assumption
- *Monte Carlo simulations*
  - Expensive (many forward runs)
  - Difficult to report
  - Full distribution
  - No distributional assumptions

25

Linear Error Propagation Analysis (First-Order-Second-Moment)

- *Assumptions*
  - Change in model prediction Δz can be approximated by a linear function of the parameter changes Δp
  - Δp is (log-)normally distributed

$$\mathbf{C}_{\hat{z}\hat{z}} = \mathbf{J}\mathbf{C}_{pp}\mathbf{J}^T$$

- Error band is *symmetric*, representing (log-) normally distributed prediction errors
- May assign certain probability to *unphysical* system behavior

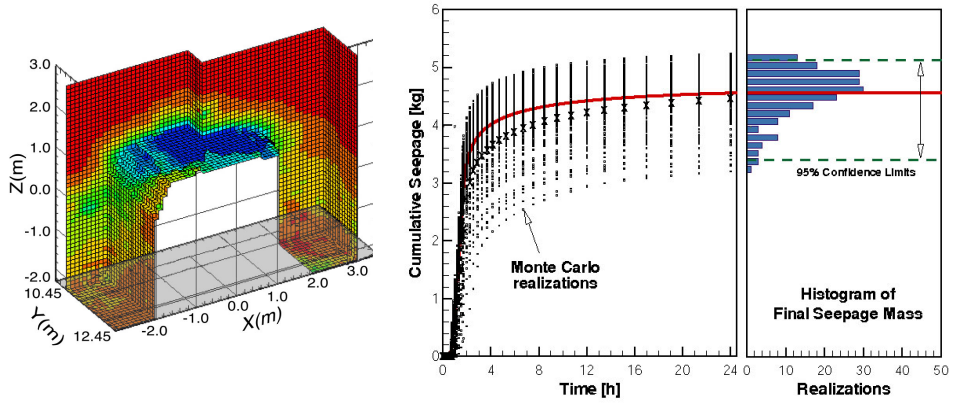
26

Monte Carlo Simulations

- Run many simulations with randomly selected parameter combinations drawn from the given probability density function.
- Provides *full distribution* of prediction uncertainty (histogram), which can be analyzed statistically.
- *Nonlinearities* are automatically taken into account.
- Results are always *physically reasonable*.
- Experimental designs: *Latin Hypercube Sampling*.
- Parameter *correlations* may be included.

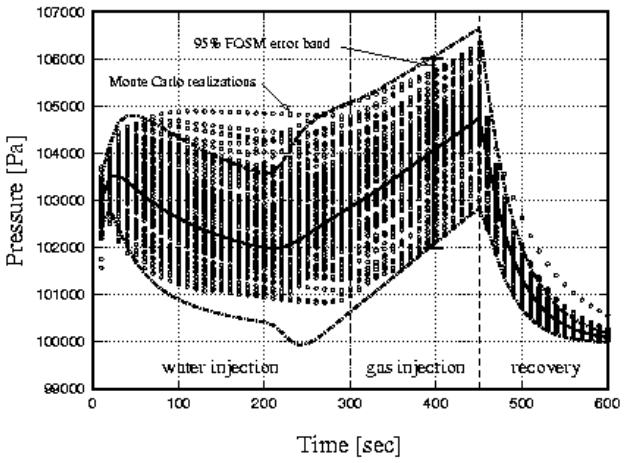
27

Monte Carlo Simulations



28

Comparison FOSM-MC

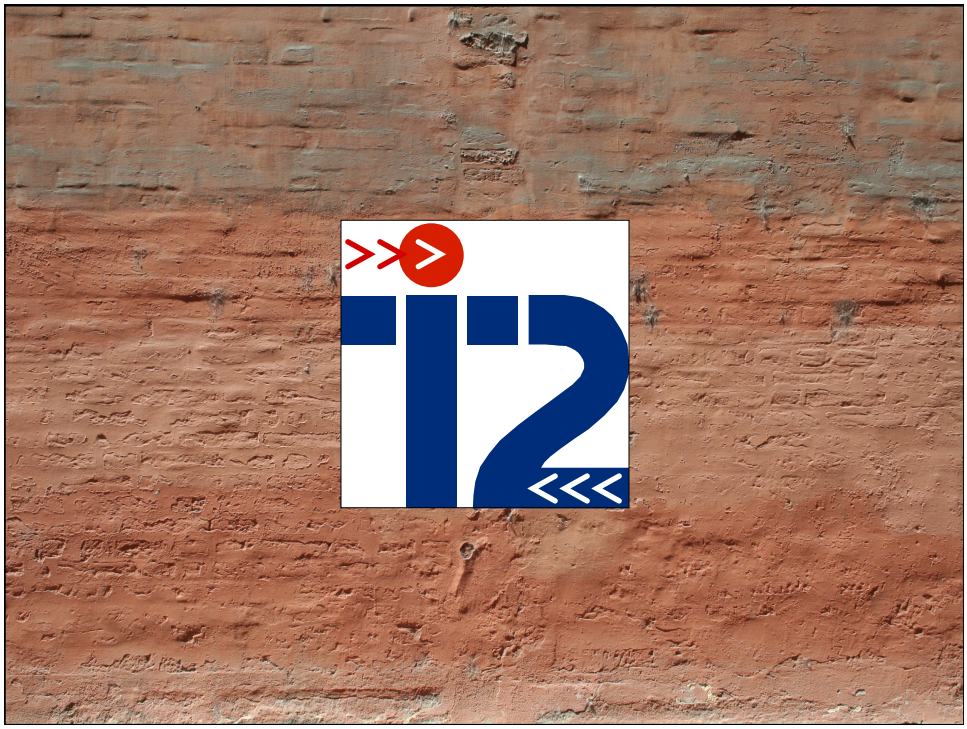


29

Uncertainty Analysis: Questions

6. Describe the main differences between FOSM and Monte Carlo simulations.
7. Assume you have to estimate the (hopefully small) probability that the TCE concentration at a drinking water well does not exceed a certain level.
   - Which uncertainty propagation analysis method would you choose?
   - Justify your choice.
   - Describe the procedure.

30



Uncertainty Propagation

Specify uncertainties in input parameters in block
> PARAMETERS or provide covariance matrix of estimated parameters using one of several formats

```
> COMPUTATION
  >> ERROR
     >>> FOSM
```

```
> COMPUTATION
  >> STOP
     >>> number of MC SIMULATIONS: 1000
     <<<
  >> ERROR
     >>> MONTE CARLO SEED: 9999 GENERATE
     >>> LATIN HYPERCUBE SAMPLING
     >>> EMPIRICAL ORTHOGONAL FUNCTION
```

32