Dimension reduction and sequential design for the emulation of complex computer models.
Application to tsunami simulation

Joakim Beck, Serge Guillas, Simon Day
University College London

Thanks:

- NERC PURE (2012-2016),
- NERC follow-on fund (2013-2014),
- NERC-DFID-ESRC Big Data for resilience (2015), NERC GCRF (2017)
- Royal Society UK-India seminar (2014) & Newton International fellowship (2016-2018)
- NERC IAA (2015)
- EPSRC IAA (2015-2016, 2016-2017)

Sequential design: Reduced number of simulations

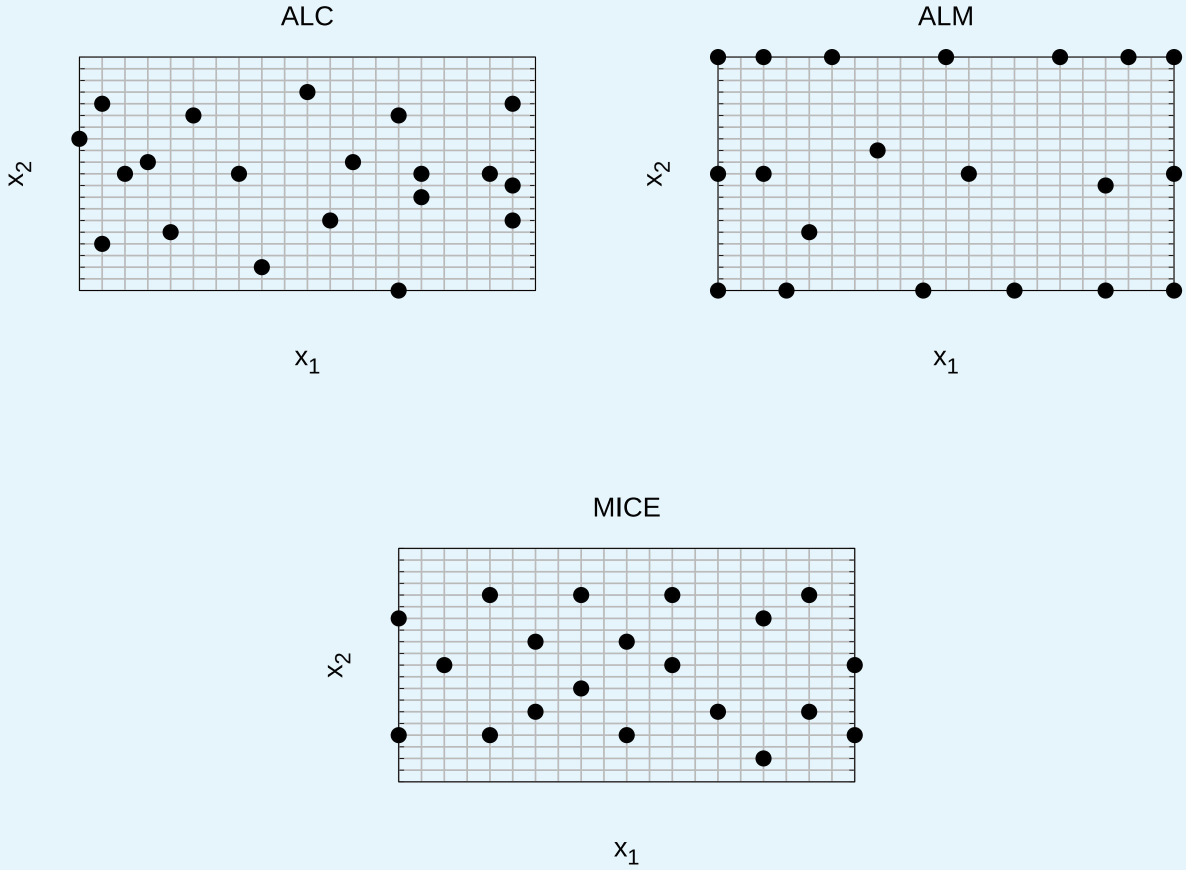
Next point(s) chosen by:

- Advanced Learning MacKay (ALM).
  High variance.
- Active Learning Cohn (ALC). Variance reduction.
  (Gramacy and Lee, 2009)
- Mutual Information for Computer Experiment (MICE): "uncertainty" at untried points with/without knowledge of selected point.

  [Beck and Guillas, 2016, SIAM/ASA J. on Uncertainty Quantification]

  Th. The adaptive algorithm is within (1-1/e) of the optimal design

Illustration



Assume $Y(x)$ GP emulator of $y(x)$ with mean $\hat{y}(x)$ and predictive variance $\hat{s}^2(x)$

ALM Algorithm

- **At stage k, ALM chooses the point $\boldsymbol{x}_{k+1}$ that max. predictive variance:**

$$\boldsymbol{x}_{k+1} = \arg\max_{\boldsymbol{x} \in \boldsymbol{X}_{cand}} \hat{s}_k^2(\boldsymbol{x}). \tag{1}$$

- **Issue: puts points on the boundary. Number of boundary points grows rapidly with dim p:**
  Grid with N^p points, ratio of boundary points to total is $(1-(1-2/N)^p)$,
  $p = 4$, $N = 10$: 0.59%; $p = 6$, $N = 10$: 0.74%.

ALC algorithm

- **ALC chooses the design point x_{k+1} that yields the largest expected reduction in predictive variance over the design space, and is defined as:**

$$\boldsymbol{x}_{k+1} = \underset{\boldsymbol{x} \in \boldsymbol{X}_{cand}}{\arg\max} \int_{\mathcal{X}} \left(\hat{s}_k^2(\boldsymbol{x}') - \hat{s}_{k \cup \boldsymbol{x}}^2(\boldsymbol{x}') \right) \mathrm{d}\boldsymbol{x}'.$$

  Appproximation of the integral over $\mathcal{X}$ over a grid of N_{ref} reference points in the design space:

$$\boldsymbol{x}_{k+1} = \underset{\boldsymbol{x} \in \boldsymbol{X}_{cand}}{\arg\max} \frac{1}{N_{ref}} \sum_{i=1}^{N_{ref}} \left(\hat{s}_k^2(\boldsymbol{x}_i) - \hat{s}_{k \cup \boldsymbol{x}}^2(\boldsymbol{x}_i) \right).$$

Mutual information

- **2 random vectors** $\bar{Y}$ **and** $\bar{Y}'$ **with marginal pdfs** $p_{\bar{Y}}(\boldsymbol{y})$ **and** $p_{\bar{Y}'}(\boldsymbol{y}')$**, and joint pdf** $p_{\bar{Y},\bar{Y}'}(\boldsymbol{y}, \boldsymbol{y}')$
- **the mutual information is equivalent to KL-divergence of** $p_{\bar{Y},\bar{Y}'}$ **and** $p_{\bar{Y}} p_{\bar{Y}'}$ **:**

$$\textbf{MI}(\bar{Y}; \bar{Y}') = \int\int \dots \int \log\left(\frac{p_{\bar{Y},\bar{Y}'}(\boldsymbol{y}, \boldsymbol{y}')}{p_{\bar{Y}}(\boldsymbol{y}) p_{\bar{Y}'}(\boldsymbol{y}')}\right) p_{\bar{Y},\bar{Y}'}(\boldsymbol{y}, \boldsymbol{y}') \, \mathrm{d}\boldsymbol{y} \, \mathrm{d}\boldsymbol{y}',$$

with $\log(0)0 = 0$**.**

- **Caselton and Zidek '84: MI to design sampling networks:**

$$\boldsymbol{X}_N^* = \arg\max_{\boldsymbol{X}_N \subset \boldsymbol{X}_{cand}} \mathbf{MI}(\bar{Y}[\boldsymbol{X}_G \backslash \boldsymbol{X}_N]; \bar{Y}[\boldsymbol{X}_N]),$$

where

- X_G **is a discrete design space,**
- $X_{cand} \subseteq X_G$ **is the set of candidate points available for selection.**
- **optimization problem is NP-hard!**

- Krause et al. 2008: alternative to avoid the need to directly solve optimization problem
- sequential algorithm that

$$\max_{\boldsymbol{x} \in \boldsymbol{X}_{cand}} \mathbf{MI}\left(\bar{Y}[\boldsymbol{X}_k \cup \boldsymbol{x}]; \bar{Y}[\boldsymbol{X}_G \backslash (\boldsymbol{X}_k \cup \boldsymbol{x})]\right) - \mathbf{MI}\left(\bar{Y}[\boldsymbol{X}_k]; \bar{Y}[\boldsymbol{X}_G \backslash \boldsymbol{X}_k]\right)$$

- can be written for GP as:

$$\arg\max_{\boldsymbol{x} \in \boldsymbol{X}_{cand}} \widehat{s}_k^2(\boldsymbol{x}) / \widehat{s}^2_{G \backslash (k \cup \boldsymbol{x})}(\boldsymbol{x})$$

Problem with MI on non equistant grids!

In computer experiments..

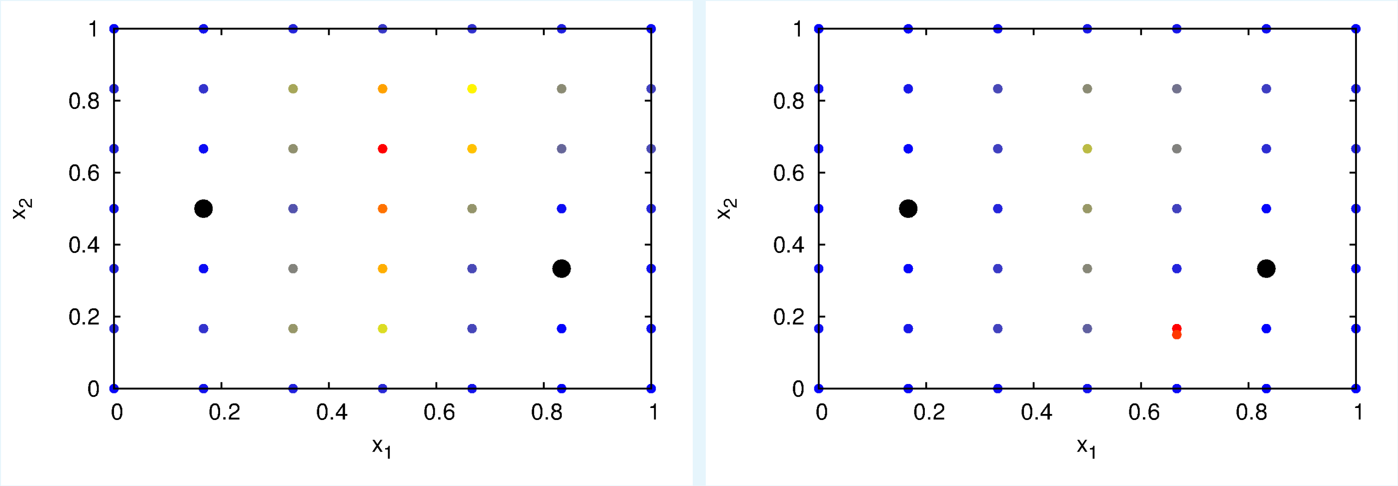


Figure 1: The score value of the MI criterion over a 7×7 equidistant grid (Left), and of the same grid with an additional point at $(2/3, 0.15)$ (Right).

Effect of nugget

Theorem [Beck and Guillas, SIAM J. UQ 2016]

For a GP emulator with constant mean, the predictive variance, at any design point $s_i \in D$ can be written as

$$\widehat{s}^2_{\tau^2}(s_i) = \sigma^2 \left(\tau^2 - \tau^4 \mathbf{e}_i^T (K + \tau^2 I)^{-1} \mathbf{e}_i + \tau^4 \frac{(\mathbf{e}_i^T (K + \tau^2 I)^{-1} \mathbf{1})^2}{\mathbf{1}^T (K + \tau^2 I)^{-1} \mathbf{1}} \right), \quad (2)$$

where $\tau^2 > 0$ is a nugget parameter, and $\mathbf{e}_i$ is the i-th unit vector.

Effect of nugget

- whenever $\tau^2 > 0$ is added to the correlation matrix diagonal, the variance of a GP at a design point consists of terms in the order of $\sigma^2\tau^2$ and $\sigma^2\tau^4$.
- in practice, the nugget τ^2 is usually orders of magnitude smaller than 1.
- In Eq. (2), magnitude last term typically much smaller than the other two;
  so predictive variance is here typically smaller than $\sigma^2\tau^2$.

- as τ^2 increases, the second and third term approaches τ^2 and τ^2/n, respectively, where n is the number of points in the design.

  This follows from: as τ^2 increases the inverse matrix reduces to $(K+\tau^2 I)^{-1} \approx \tau^{-2} I$.

  Hence, if τ^2 is large enough, we can show by a simple calculation using Theorem that $\widehat{s}^2_{\tau^2}(s_i) \approx \sigma^2\tau^2/n$ for $s_i \in D$.

MICE criterion

- **MICE**

$$\boldsymbol{x}_{k+1} = \underset{\boldsymbol{x} \in \boldsymbol{X}_{cand}}{\arg\max}\, \hat{s}_k^2(\boldsymbol{x}) / \hat{s}^2_{G \setminus (k \cup \boldsymbol{x})}(\boldsymbol{x}; \tau_s^2), \tag{3}$$

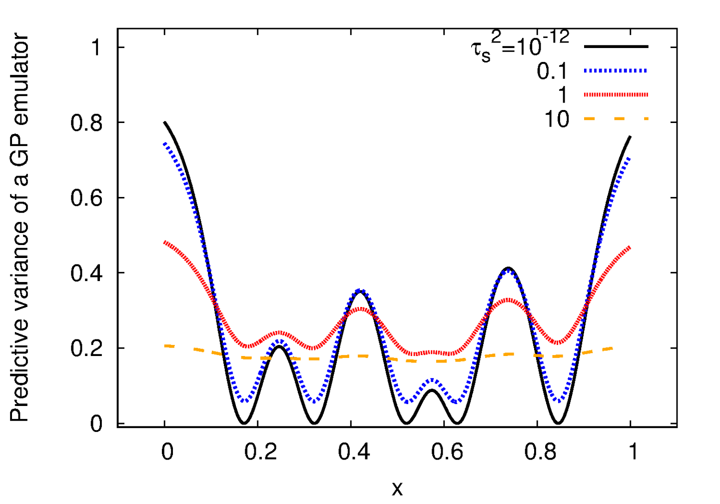
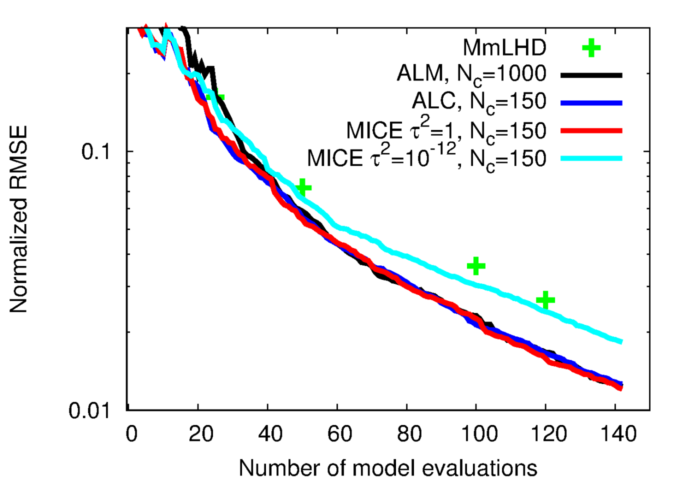
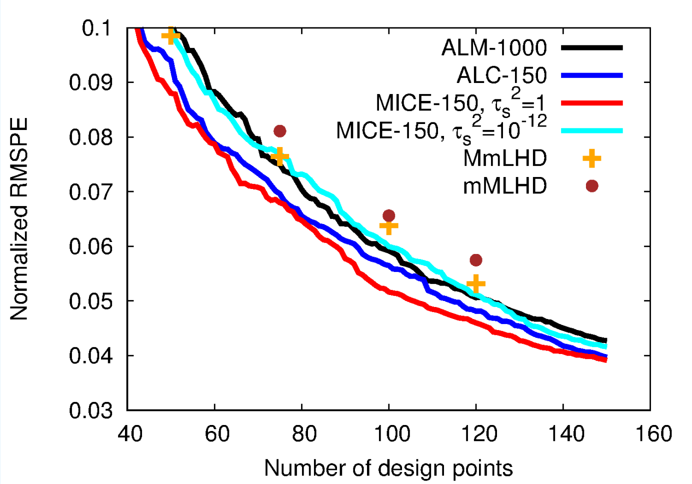


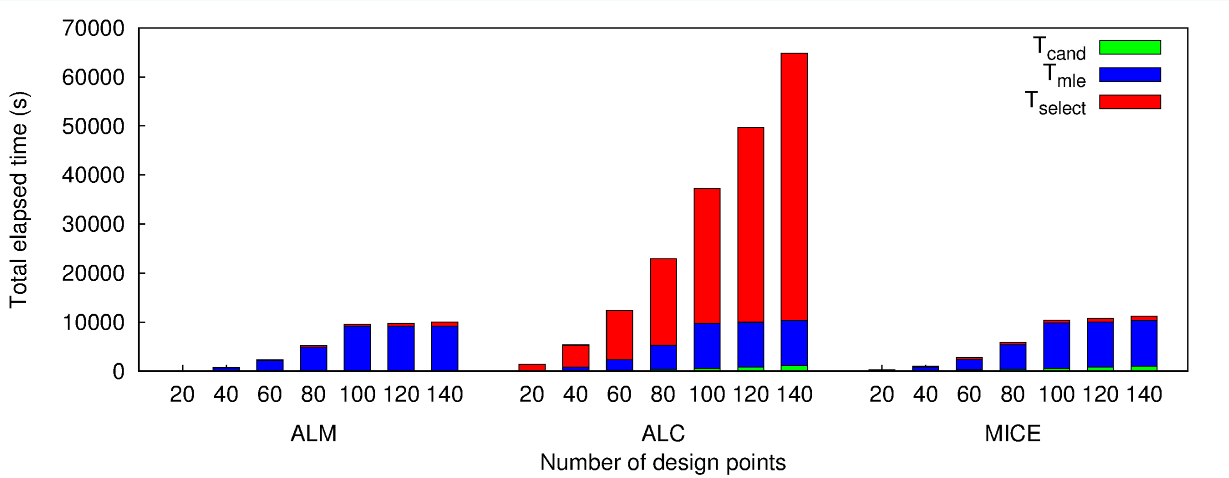
Figure 2: The predictive variance of a GP emulator as a function of τ_s^2 for a one-dimensional problem in domain $[0, 1]$.

- **correlation parameters ξ are adaptively learned by MLE.**

Performance:
For Genz oscillatory 4D & 8D

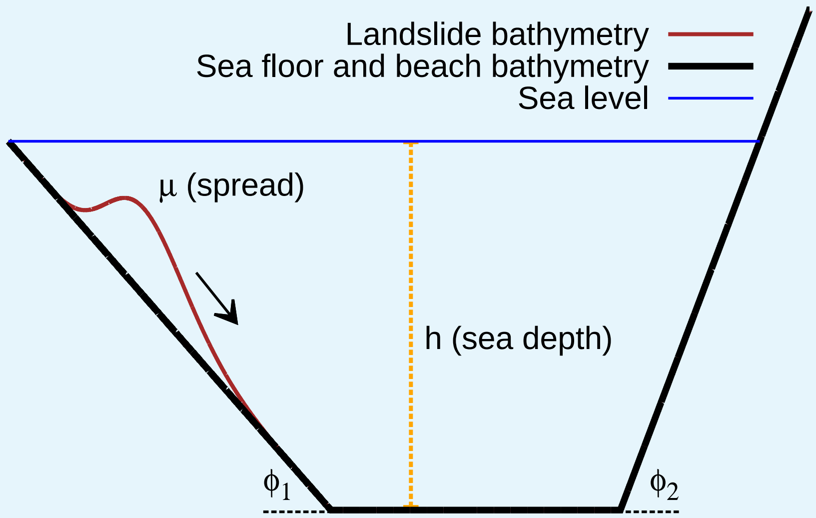






Tsunami model VOLNA: example

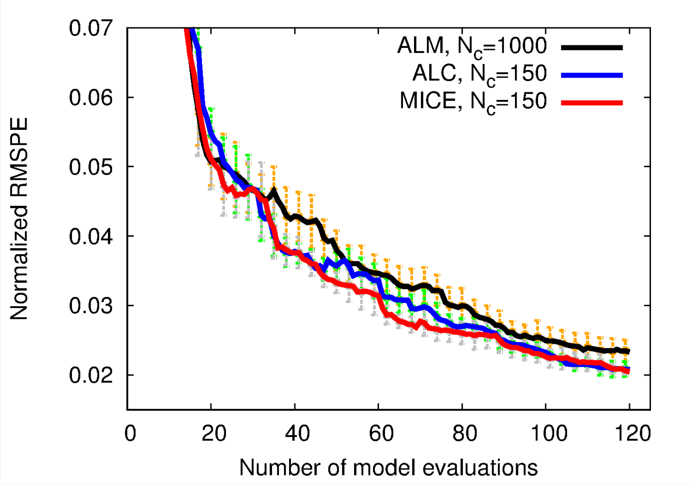
Benchmark 3 of 2004 Catalina meeting, USA.

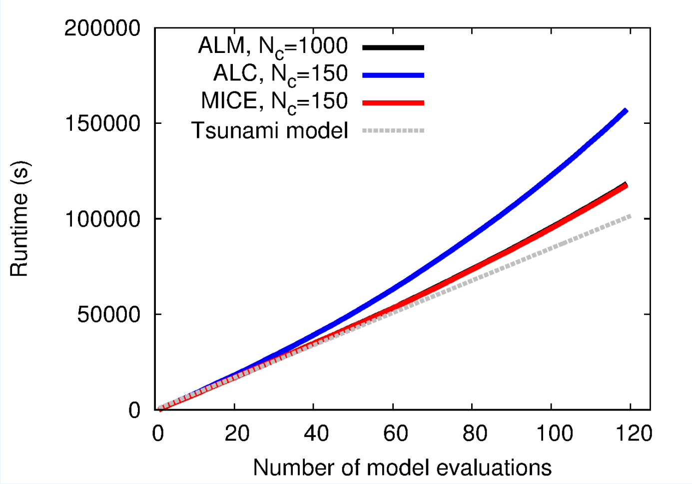


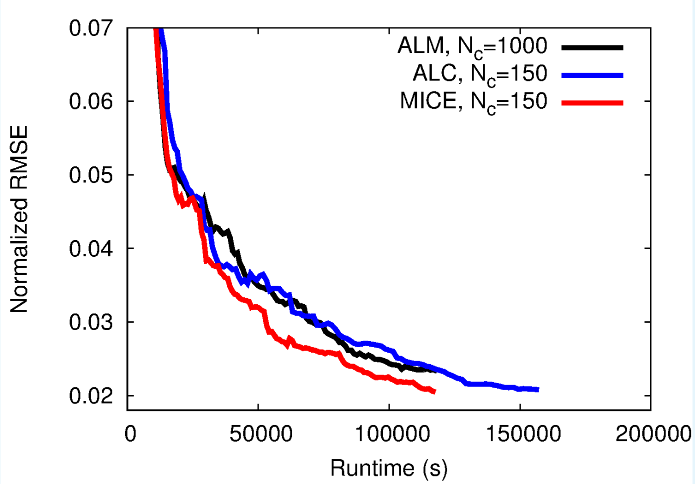
Set up

- We measure the max. height at x=2800m
- Simulation time T=170 seconds. Wall-clock time, 1 GPU: 5-10 min.
- 500 maximin LHD points for the comparison.
- Max elevation range: 0.2 m to 9.5 m
- start: 5 sets of 10 random initial points for confidence bands.
- after 60 points same correlation length.

Performance:

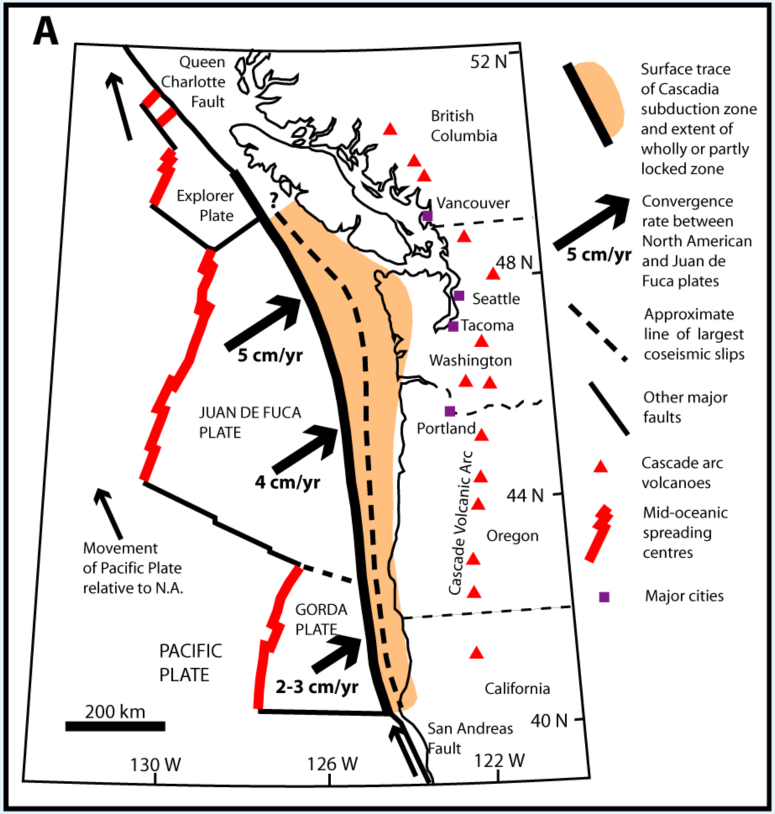






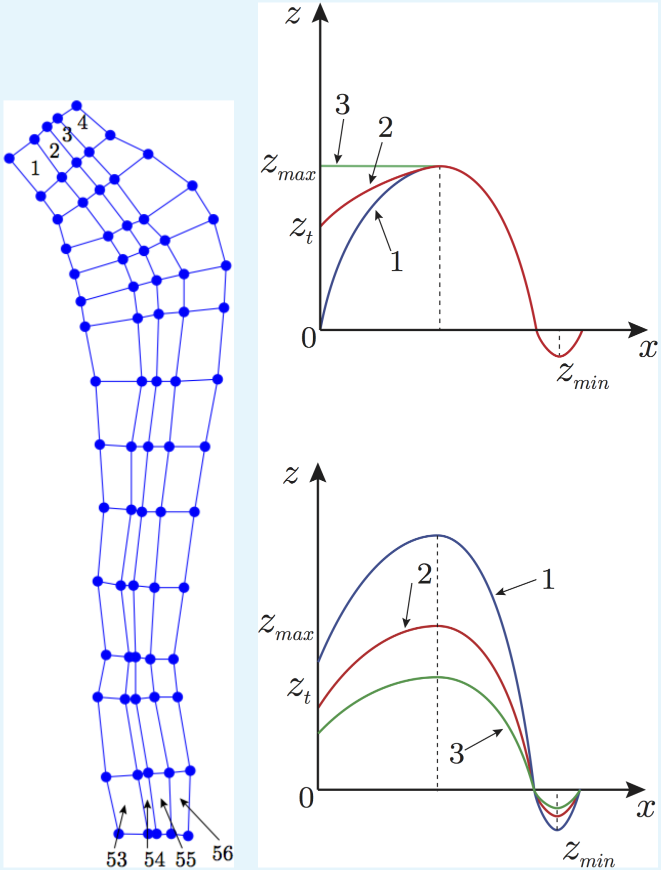
Cascadia tsunami risk

Fault system



A simple tectonic map of the Cascadia subduction zone

Uplift: 3 parameters



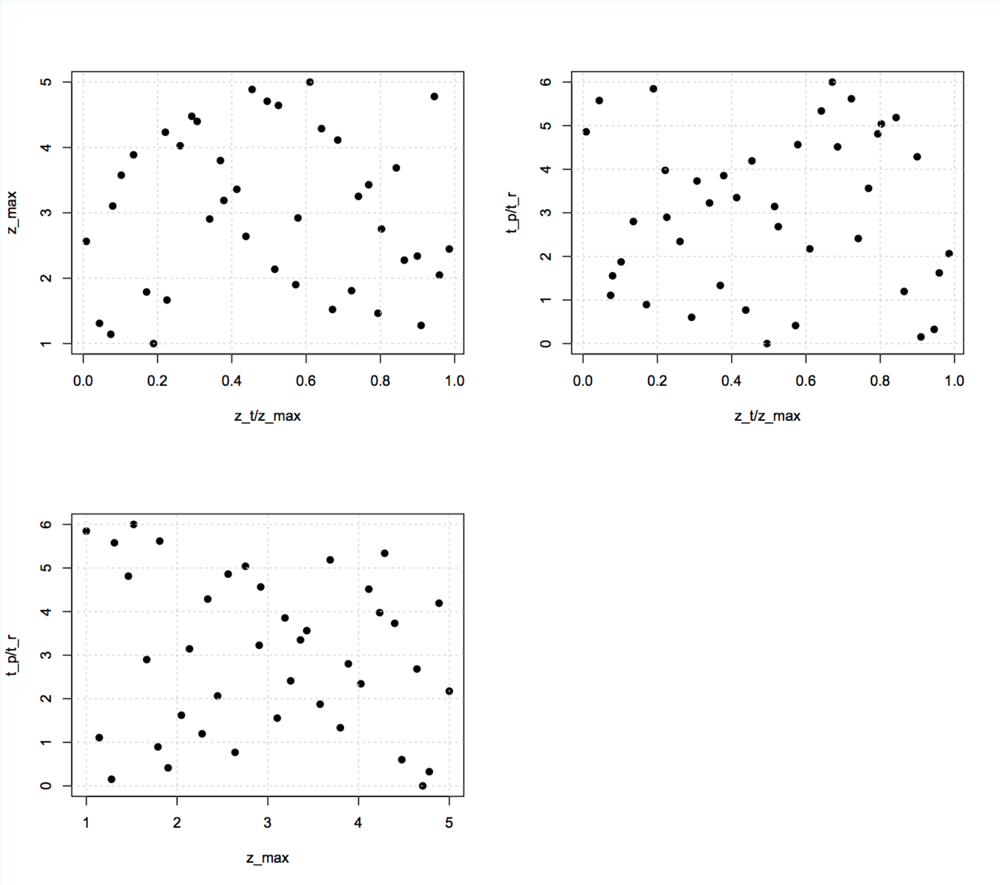
and relative time of propagation N→ S

- real seabed deformation more complex than simple elastic deformation around quake rupture
- generation mechanism: not simultaneous uplift! Uncertainties resulting from this!
- 2 left polygons rise and the rest 2 (right) subside.
- motion North to South.
- movie:

$$(z_{edge}/z_{max}, z_{max}, t_p/t_r) = (0.9, 4.8, 0.3).$$

Tsunami simulations

1. Latin Hypercube Design set of 40 runs.

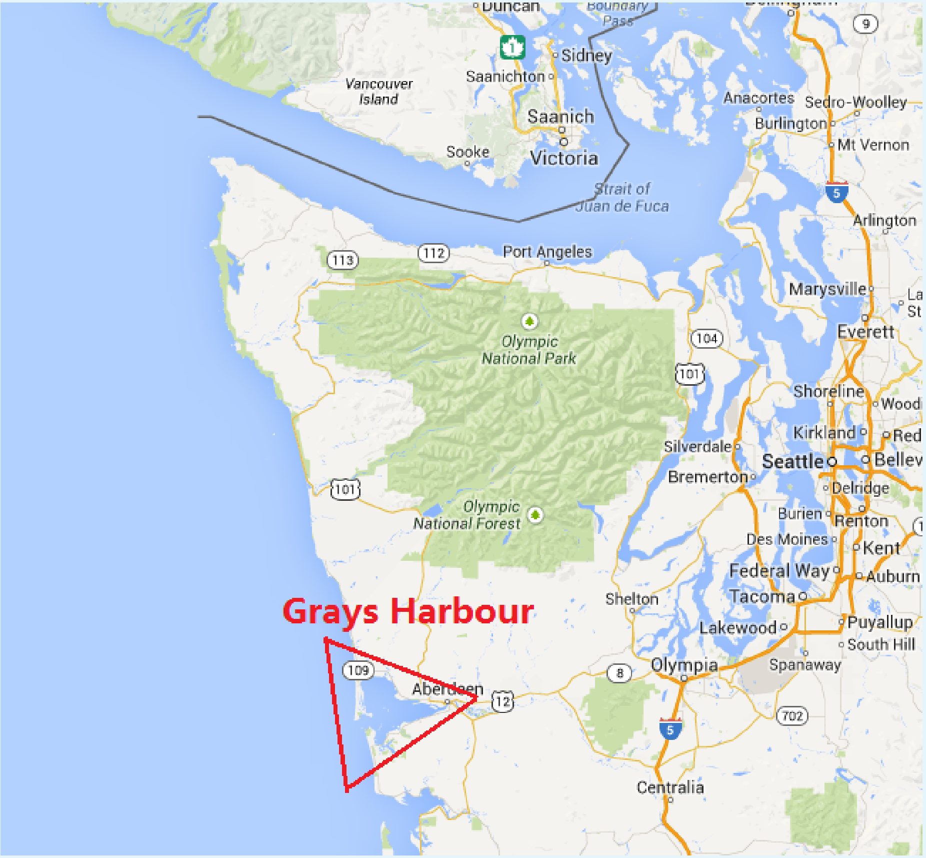


2. tsunami model VOLNA

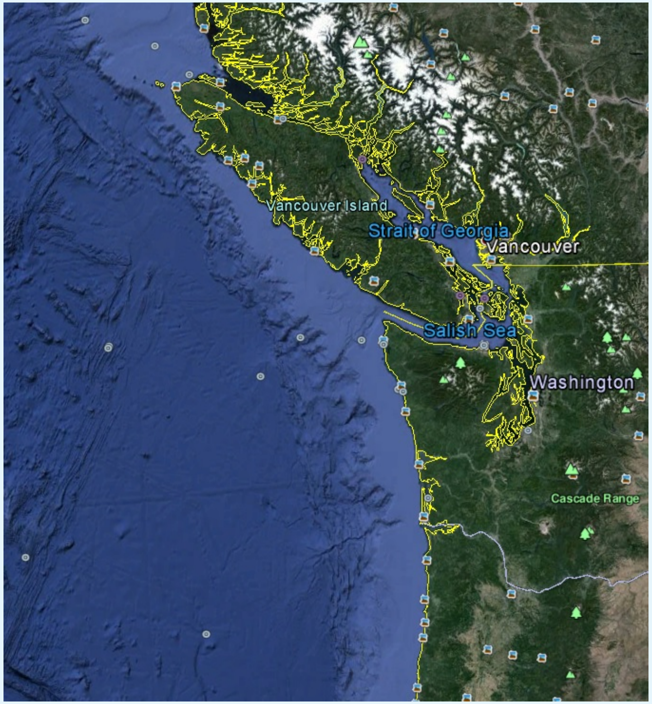
   - unstructured mesh
   - finite volumes
   - generation, propagation, inundation
   - NO friction, wave dispersion and breaking.

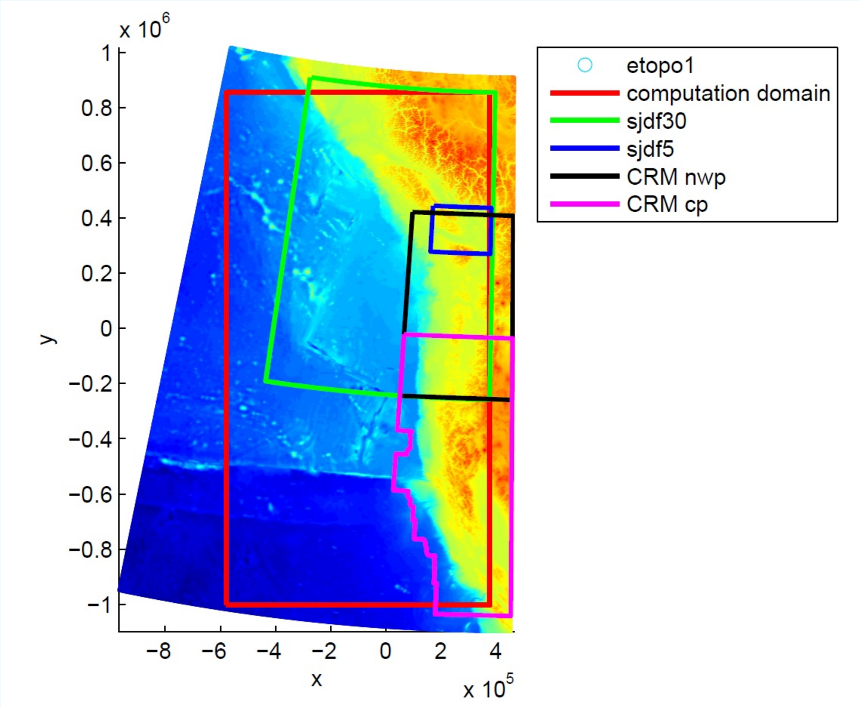
3. on Emerald GPU cluster (372 M2090 GPUs!): acceleration 3000-8000x compared to 1 CPU (István Reguly & M. Giles, Oxford)

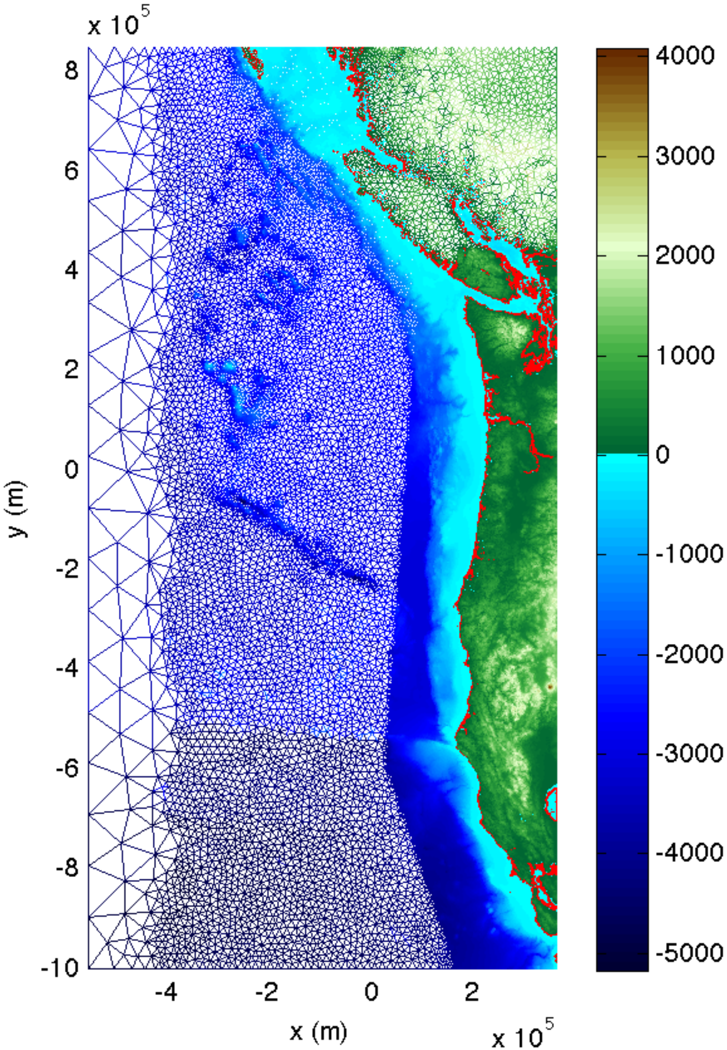
DEM information

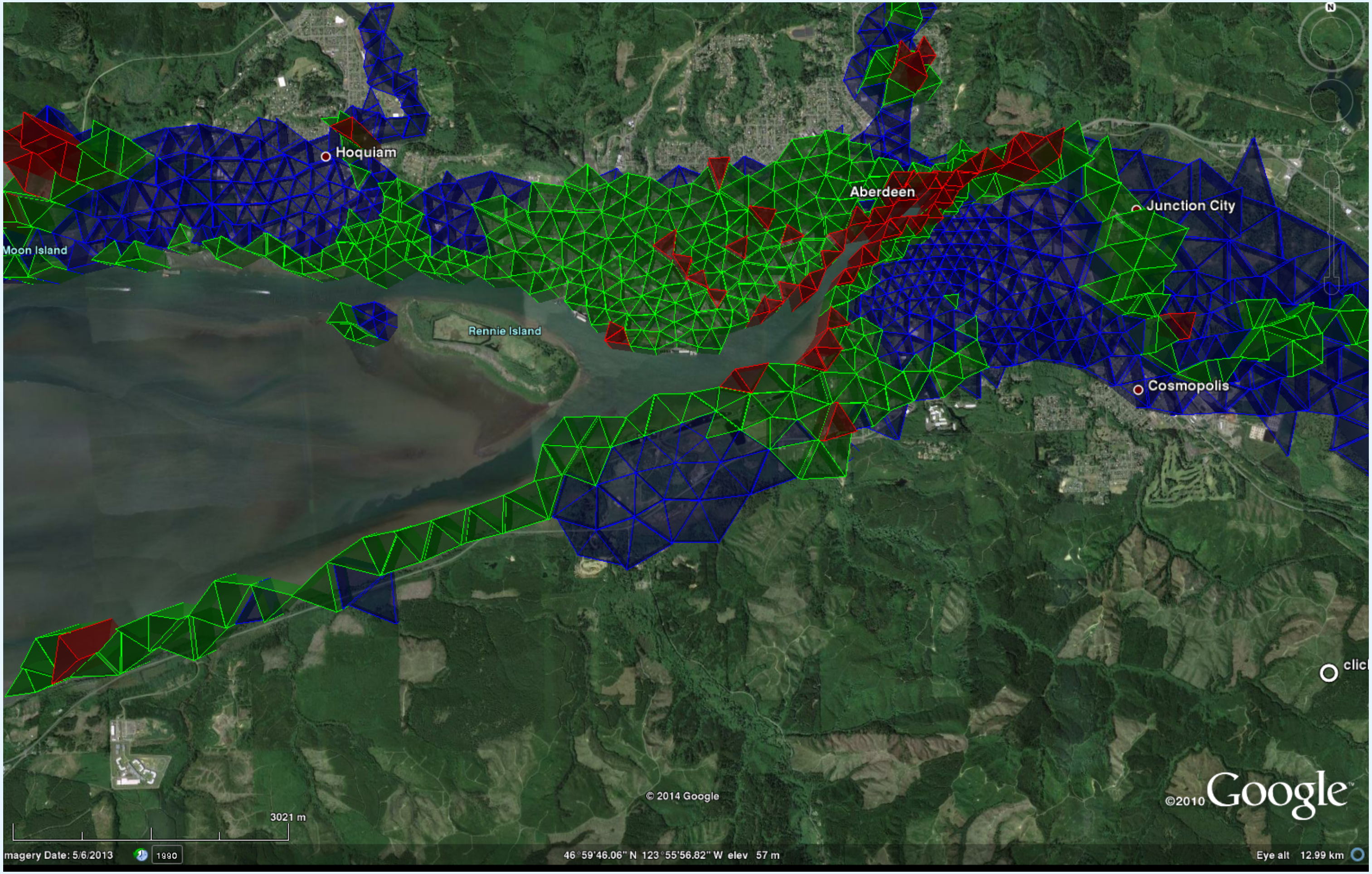


DEM information: source NOAA NGDC






x 10^5
8
6
4
2
0
-2
-4
-6
-8
-10
y (m)
-4
-2
0
2
x (m)
x 10^5
4000
3000
2000
1000
0
-1000
-2000
-3000
-4000
-5000

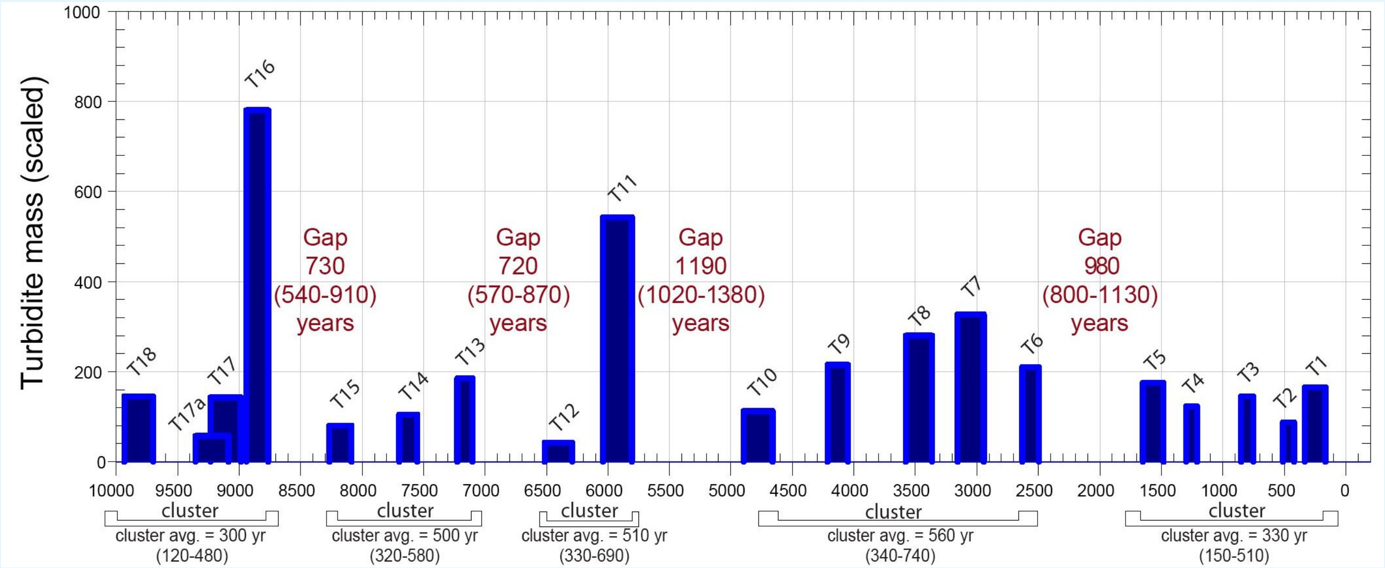


Hoquiam
Aberdeen
Junction City
Moon Island
Rennie Island
Cosmopolis
© 2014 Google
©2010 Google
3021 m
magery Date: 5/6/2013
1990
46°59'46.06" N 123°55'56.82" W elev 57 m
Eye alt 12.99 km

©2010 Google

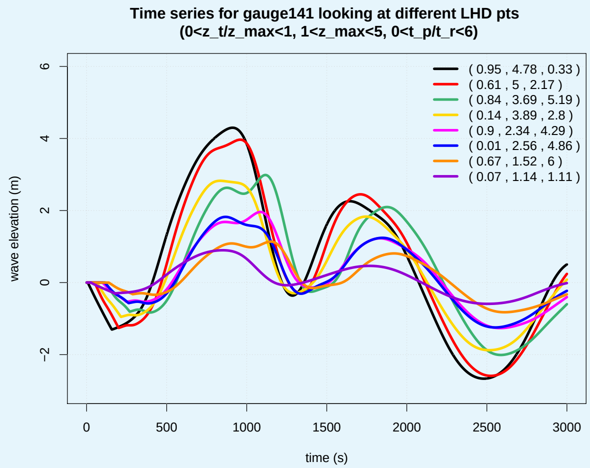
Time: 2610,025
Inun
29,3
20
10
0
-10
-17,

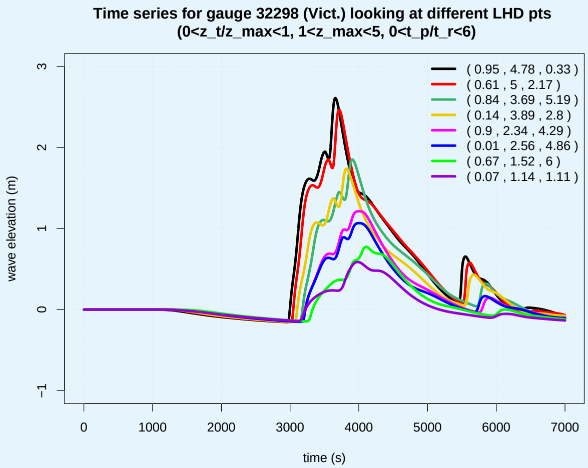
Tsunami catalogue



- actually earthquake shaking catalogue
- clustering.. important?

Various gauges: locations, variations





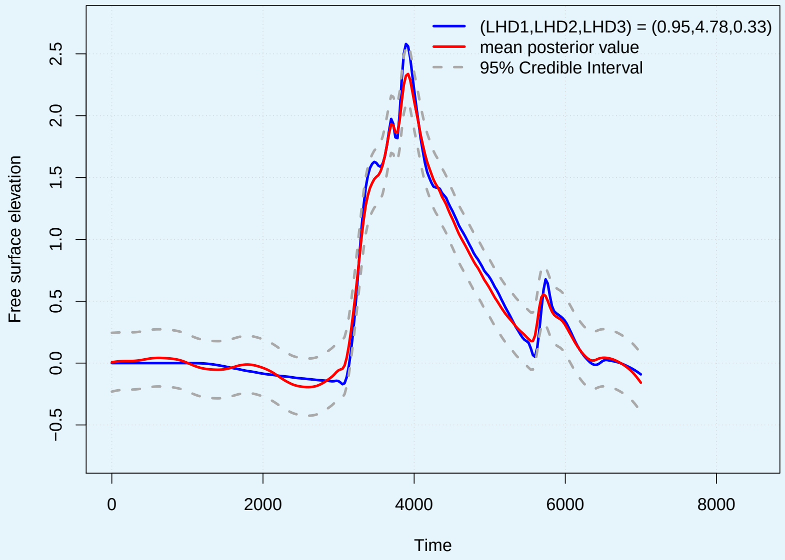
Emulation of entire time series

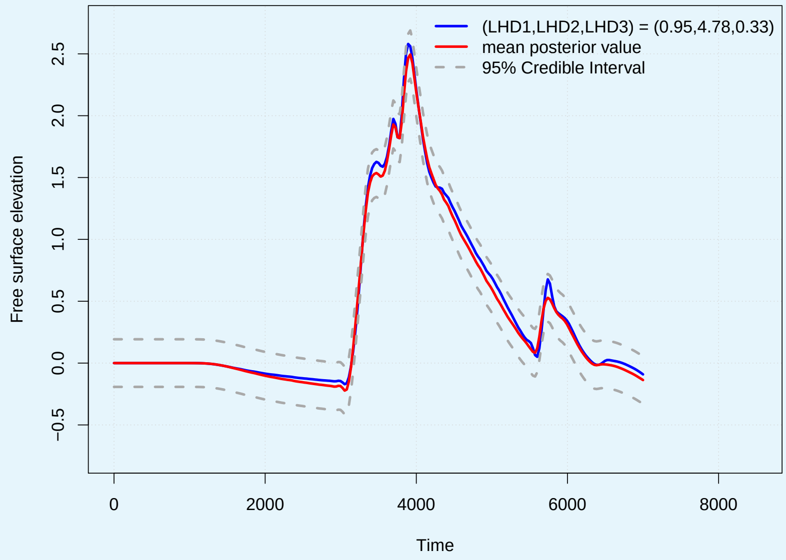
Sarri, Guillas, Dias, NHESS 2012

Sarri, Guillas, Day, Liu, Dias, 2016 (submitted)

- OPE emulator (speed & dimension)
- Polynomials for inputs regressors
- Fourier terms for outputs regressors
- new Functional Principal Components for outputs regressors

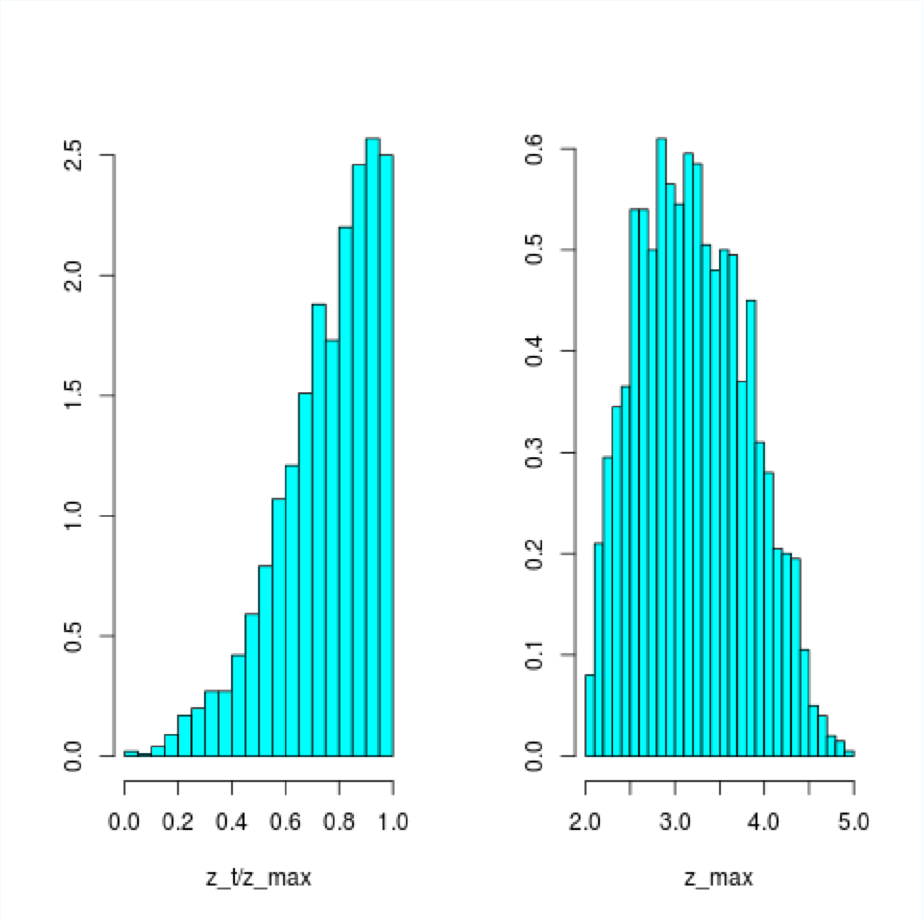
LOO diagnostics plots, gauge 32298 (LHD point 5).

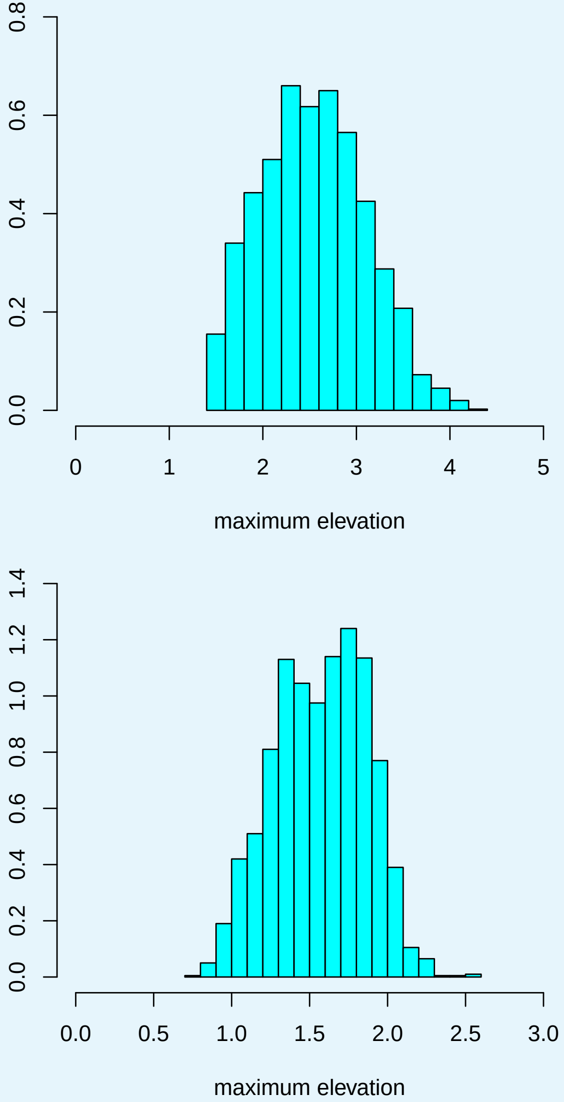




Fourier OPE v. PCs OPE

For Cascadia: propagation of only 2 parameters





gauges 141 & 32298

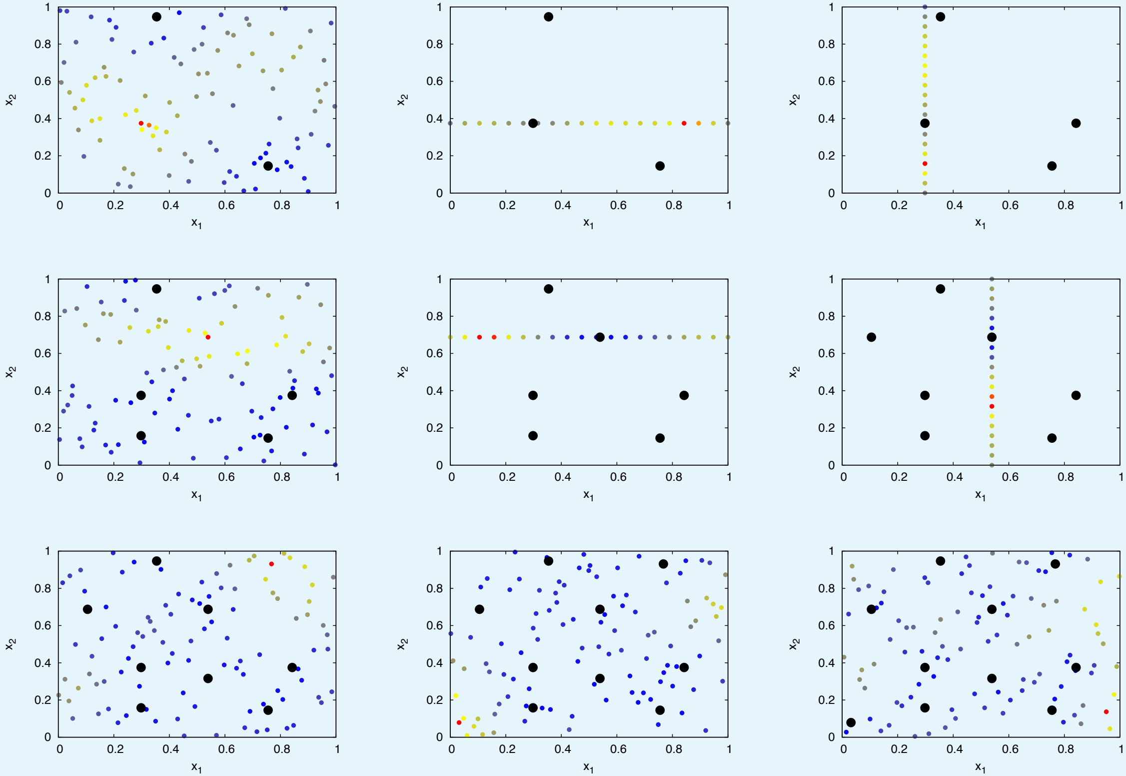
Sequential design MICE interwoven with variable selection: “Screening MICE”

- **Input variable screening used to identify parameters with negligible or linear effect**
- **Aim: reduce dimension to allow thorough exploration of the input space with less runs**
- **Combine a sequential Morris method (Boukouvalas et al., 2014) with MICE.**
- **Algorithm: first, all input variables are set to status inactive.**
- **Instead of a fixed perturbation step size for each direction (standard sequential Morris method): MICE identifies the most informative design point along the direction.**

Screening MICE: Algorithm

1. Perform one iteration of MICE to select a new point $\mathbf{x}^*$
2. Select a new point with MICE for each still inactive input variable
3. Calculate/update the (Morris) moment estimates μ_i and σ_i using the $1 + p_{inactive}$ function evaluations
4. Set variable i to (nonlinear) active if $\sigma_i > \sigma_0$. Repeat from step 1, R times.
5. The dimensionality of the problem is reduced by removing inactive variables, and instead account for their linear or negligible effects in the regression term.

Screening MICE: 2D illustration



Numerical examples

- We use $R = 2$ in the sequential Morris method.
- Function:

$$y(\mathbf{x}) = \frac{1}{1 + \mathbf{cx}},$$

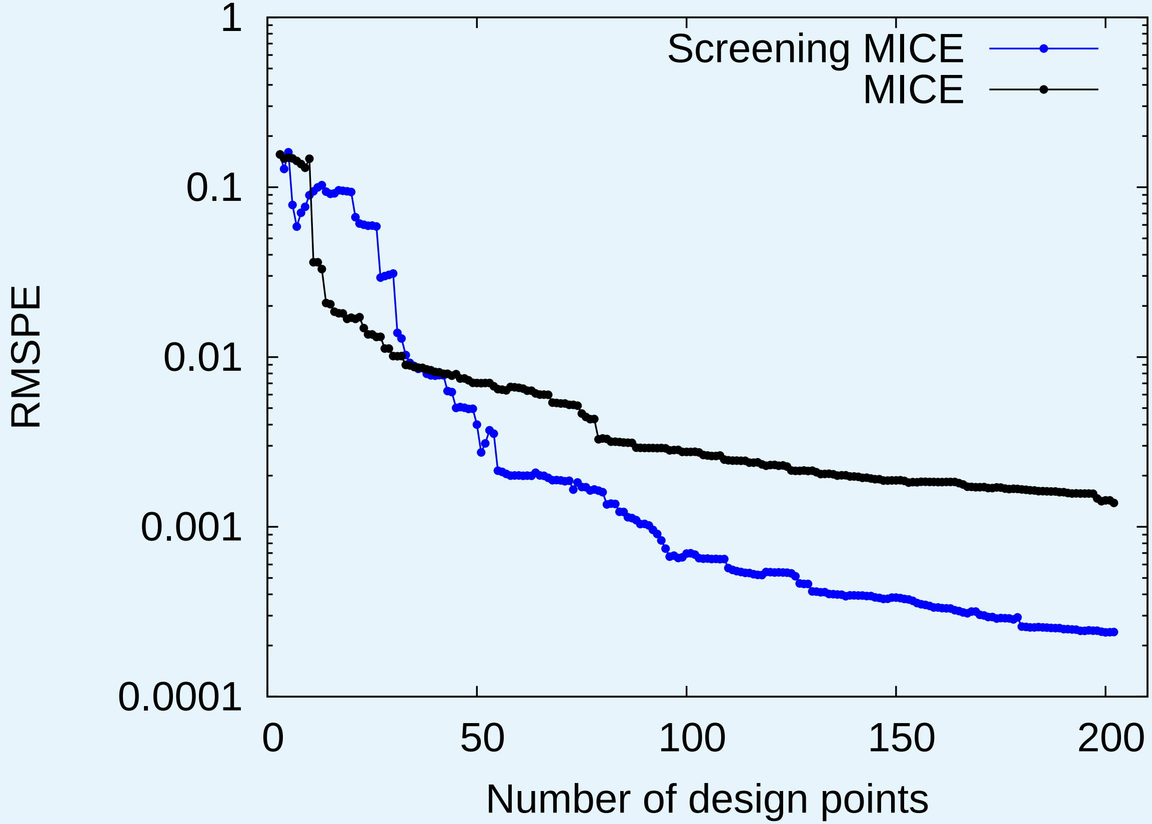
where

a.) $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ are the parameters

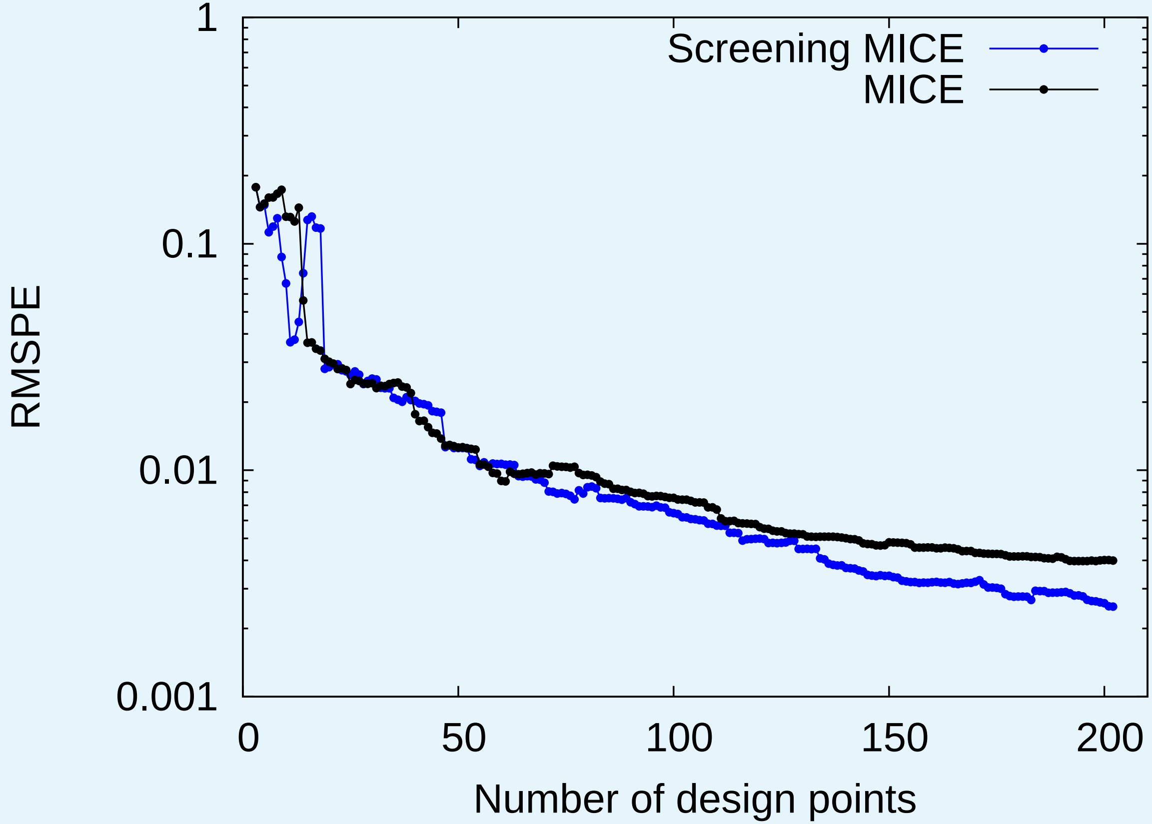
b.) $\mathbf{c} = (c_1, c_2, \ldots, c_d)$ are coefficients that determines the influence of the parameters on the output $y(\mathbf{x})$.

- Coefficients are for each problem chosen so that we can specify the number of active parameters.
- For each problem, the Morris method successfully identified all of the active parameters.

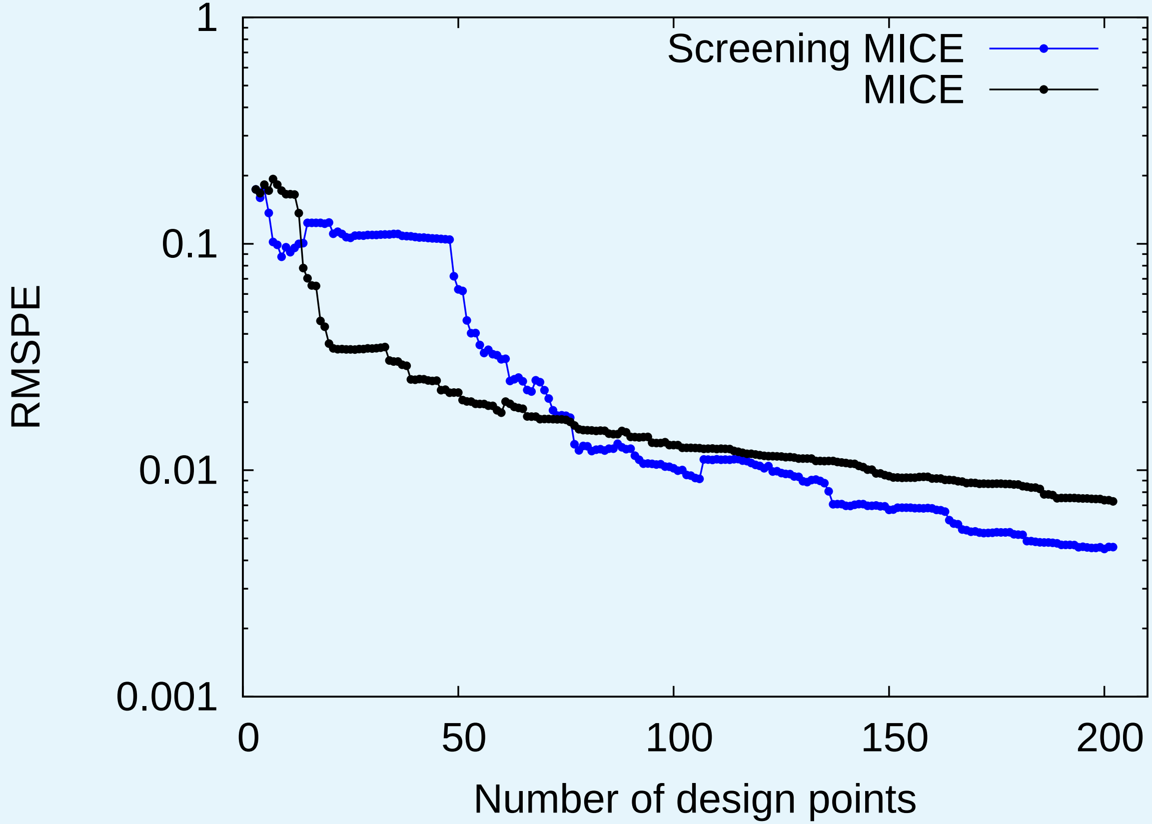
Example 10D: 4 active variables



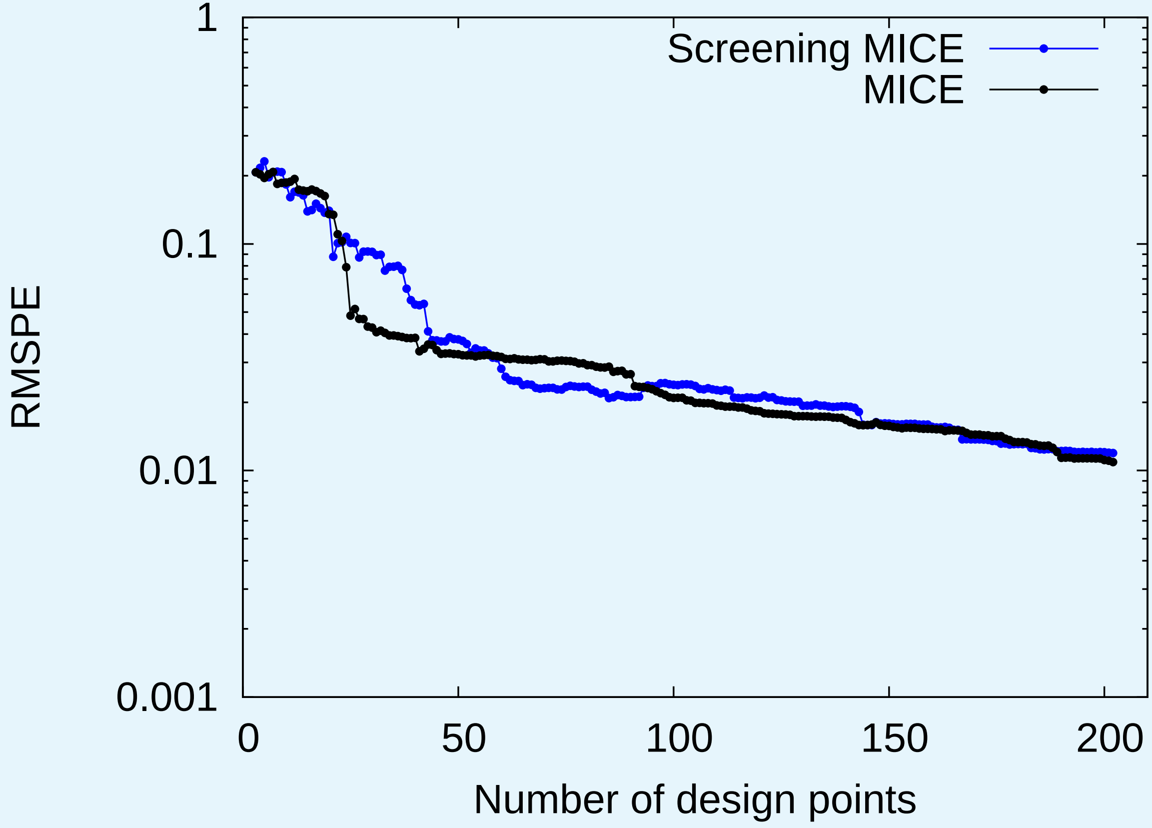
Example 10D: 8 active variables



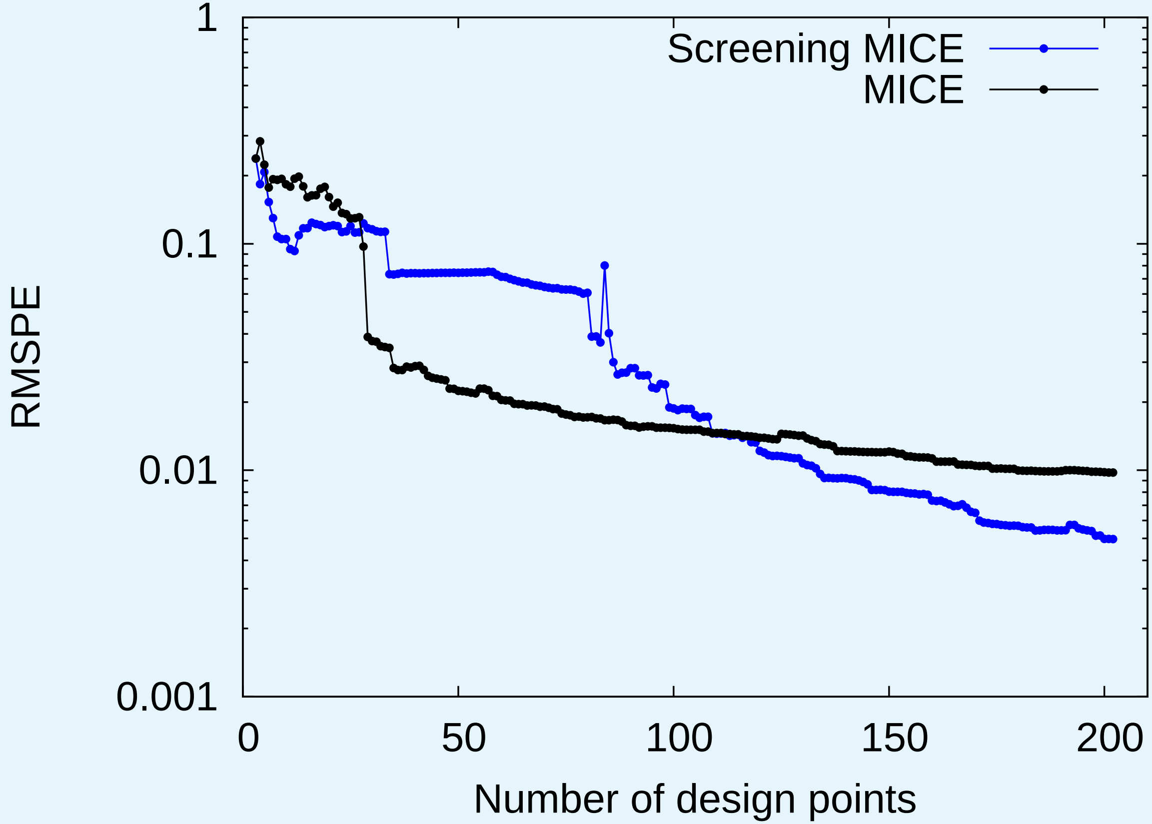
Example 20D: 8 active variables



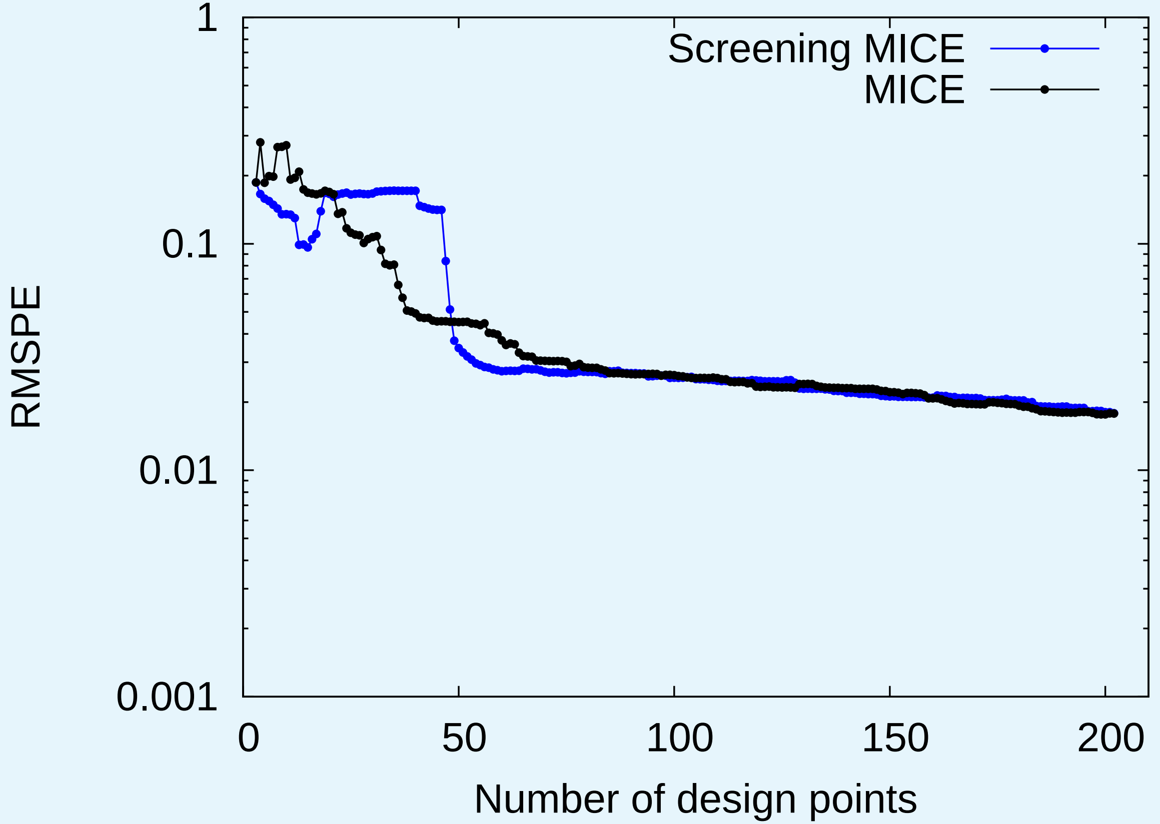
Example 20D: 16 active variables



Example 30D: 8 active variables



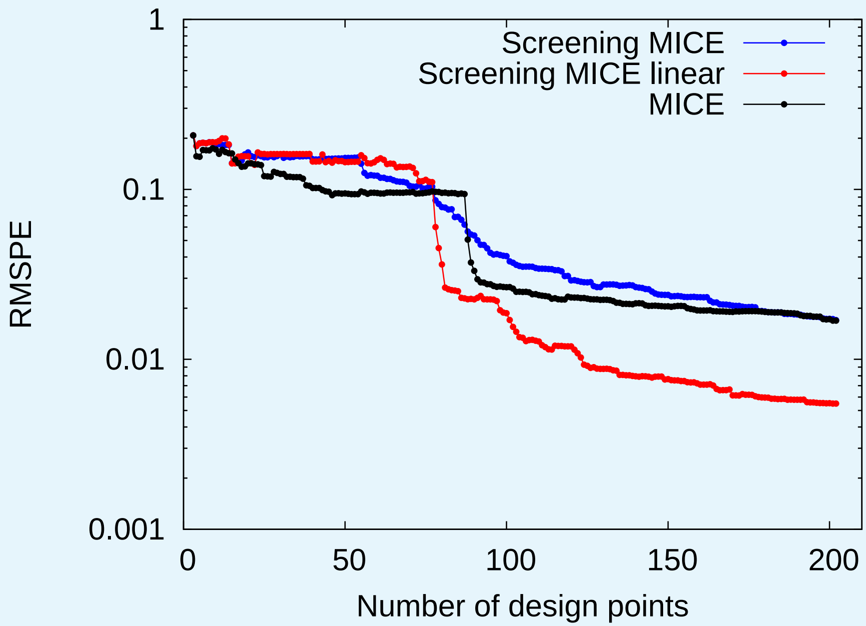
Example 30D: 24 active variables



Example 20D Welsh: linear terms

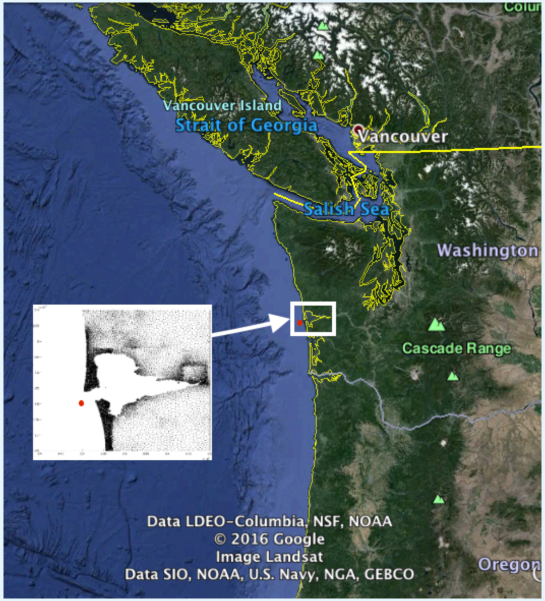
Note: Virtual Library of Simulation Experiments: Test Functions and Datasets, SFU (D. Bingham)
`http://www.sfu.ca/ ssurjano/index.html`

The screening method can identify linear terms which can improve the performance substantially:



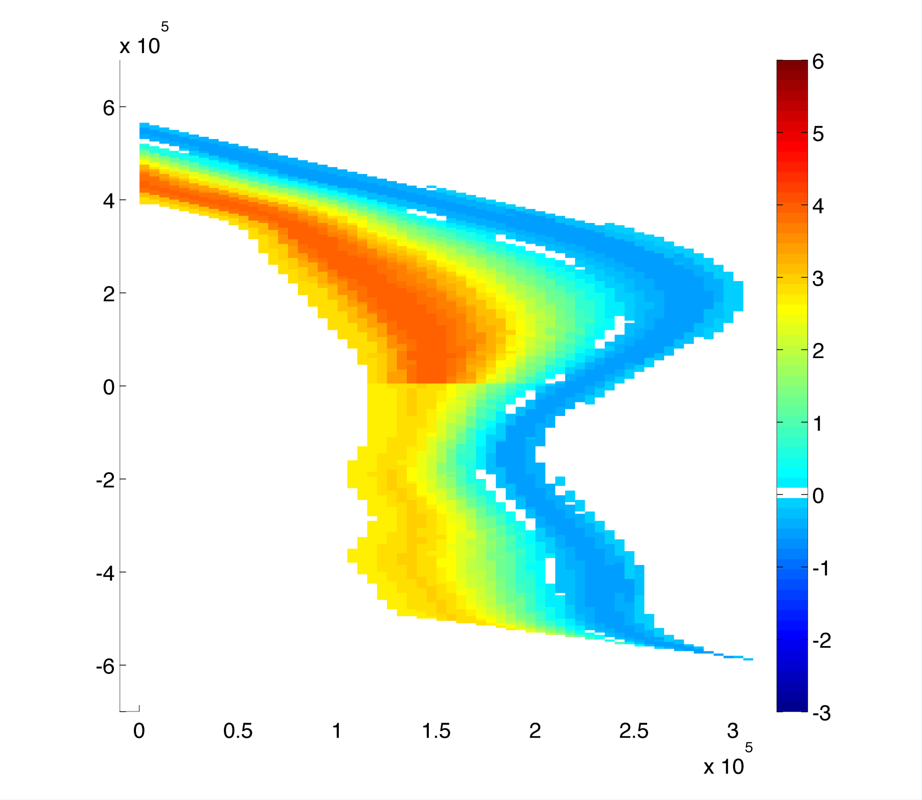
Cascadia subduction zone study

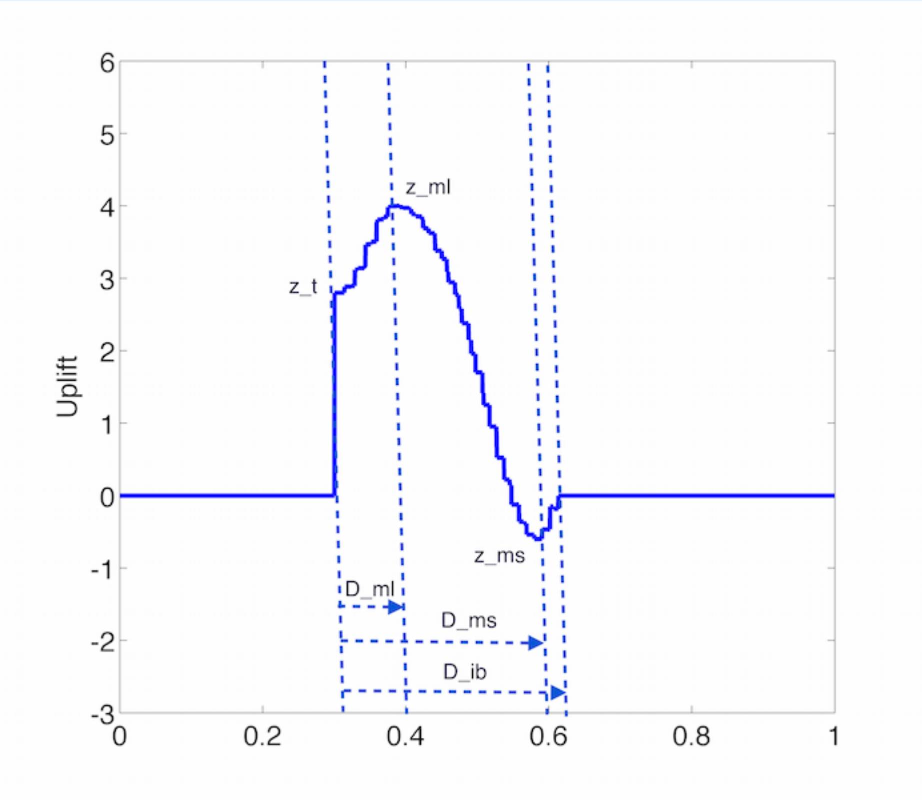
A gauge is placed outside of Grays Harbor as shown in figure below:



The red dot shows the location of the gauge. An uplift model is proposed with 9 uncertain parameters.

Uplift model parameterization





Uplift
z_ml
z_t
z_ms
D_ml
D_ms
D_ib
6
5
4
3
2
1
0
-1
-2
-3
0
0.2
0.4
0.6
0.8
1

9 uncertain parameters

t	Rupture time	$[150, 300]$	s
D_{ml}	Distance to maximum subsidence line	$[0.1, 0.3]D_{ms}$	m
D_{ib}	Distance to inboard line	$[1.1, 1.4]D_{ms}$	m
z_{ms}	Maximum subsidence uplift amount	$-[0.1, 0.3]z_{ml}$	m
$z_{ml,N}$	Maximum locking uplift amount (North)	$[2, 7]$	m
$z_{ml,S}/z_{ml,N}$	Maximum locking uplift amount in North by in South	$[0.3, 0.8]NS_r * (b+1)$	m
$z_{t,N}/z_{ml,N}$	Trench line by maximum locking line uplift amount (North)	$[0.3, 1.4]$	m
$z_{t,S}/z_{ml,S}$	Trench line by maximum locking line uplift amount (South)	$[0.3, 1.4]$	m
NS_y	North-South point	$[-110000, 60000]$	m

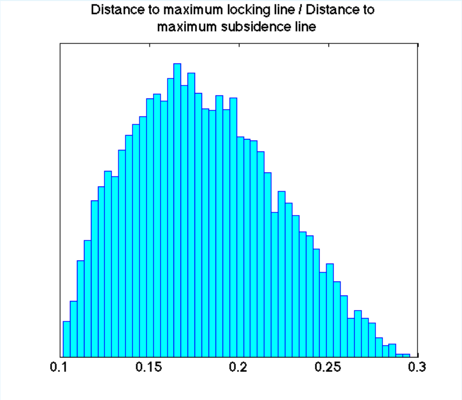
Statistical emulation

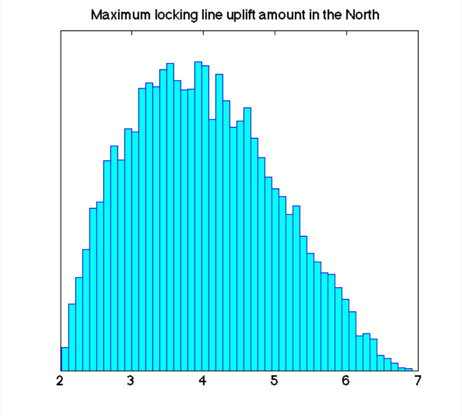
- A GP emulator is built over the 9 parameters within given ranges, starting on two random design points.
- After selecting a design with 22 input configurations with Screening MICE ($R = 3$), 3 parameters identified:
  1. D_{ib}, $z_{ml,S}/z_{ml,N}$, and $z_{t,S}$, as negligible
  2. NS_y as linear
  3. remaining parameters as nonlinear.
- Algorithm then reduces the design space by removing the negligible parameters, and adding a linear regression term for NS_y.
- Algorithm is stopped after collecting 40 design runs, and each run takes about 90 min on 3 GPUs.

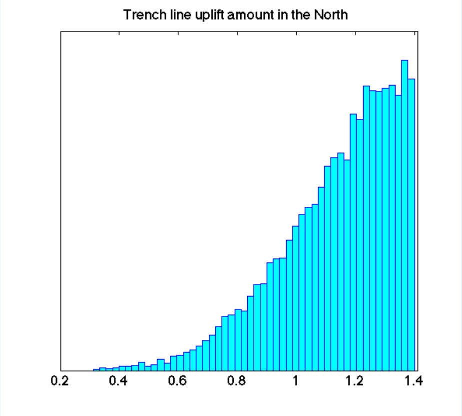
Propagation of parameter uncertainties

10,000 samples of the GP using

- $t \sim \text{Uni}([150, 300])$,
- $D_{ml} \sim \text{Uni}([0.1, 0.3]D_{ms})$,
- $z_{ms}/z_{ml} \sim \text{Beta}(2, 3, -[0.1, 0.3])$,
- $z_{ml,N} \sim \text{Beta}(2, 3, [2, 7])$,
- $z_{t,N}/z_{ml,N} \sim \text{TruncNormal}(1.4, 0.09, [0.3, 1.4])$
- $NS_y \sim \text{Uni}([-110000, 60000])$

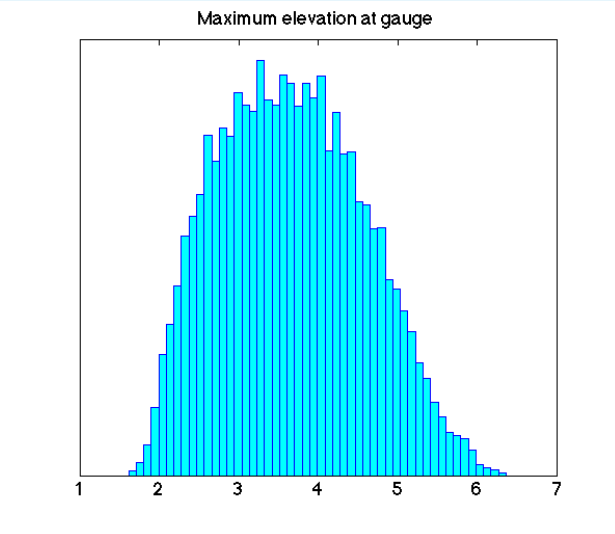




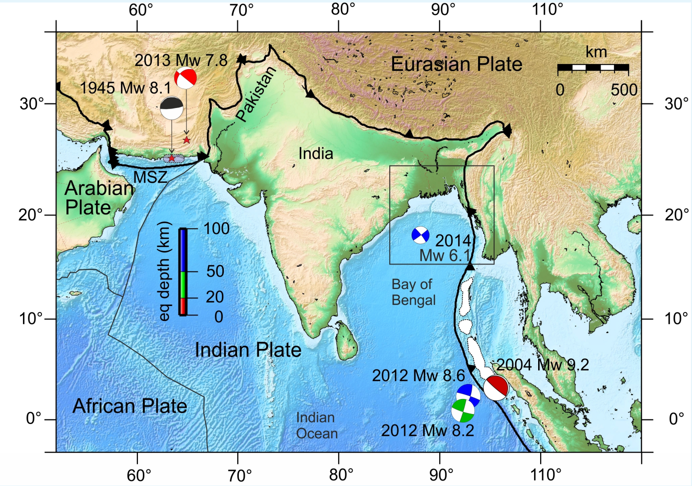


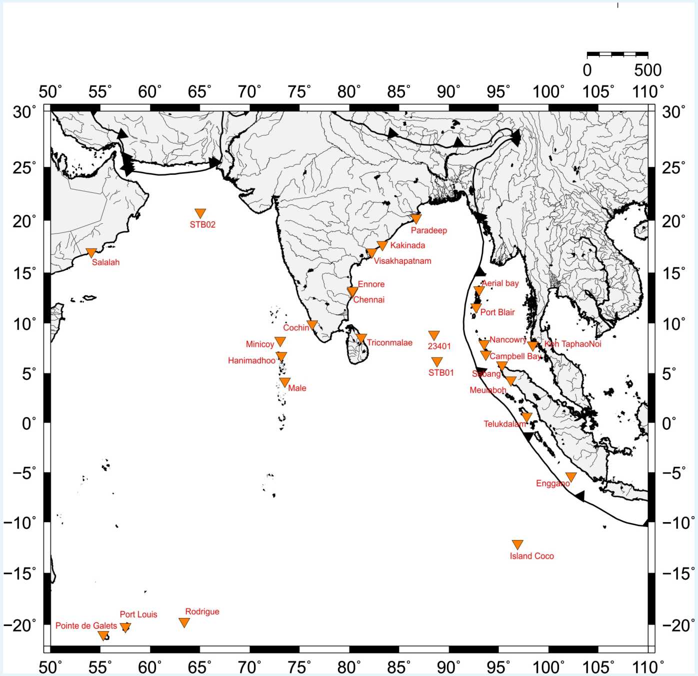
Results: pdf for the maximum elevation

At the gauge located at Grays Harbor:



Other application to Indian Ocean tsunamis: early warnings (in progress)





Source of data

- Teleseismic information
- GPS
- Ground level changes
- Satellite overpass
- Tide gauges