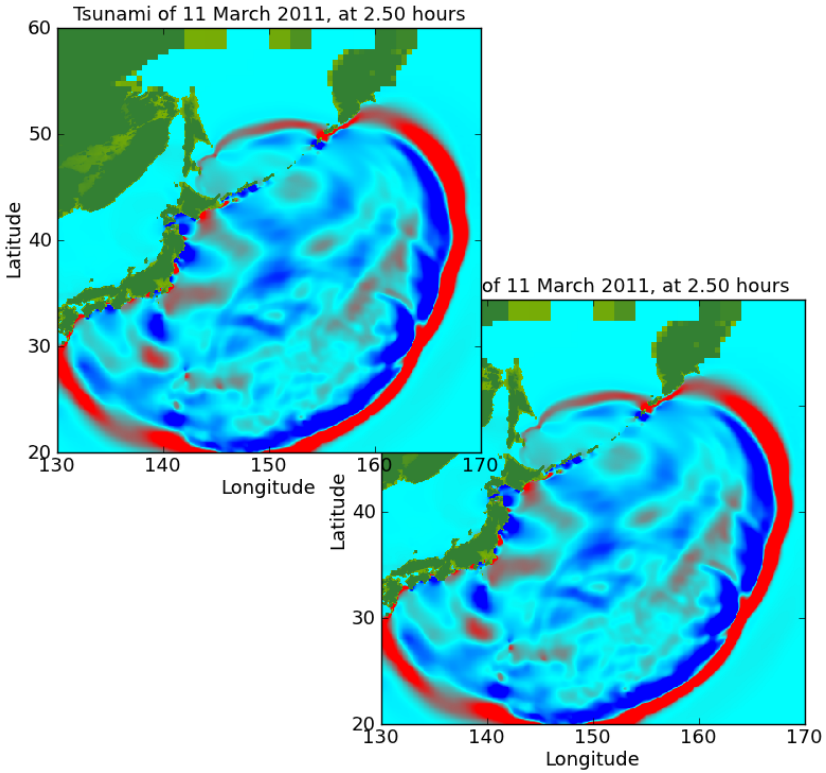
Reproducibility and Open Science

Randall J. LeVeque
Applied Mathematics
University of Washington



My main research interests

- Numerical methods for wave propagation,
  Computational Fluid Dynamics

- Since 1994: Open source Clawpack software
  (Shock waves, seismic, volcanic flows, astrophysics,
  traumatic brain injury, etc.)

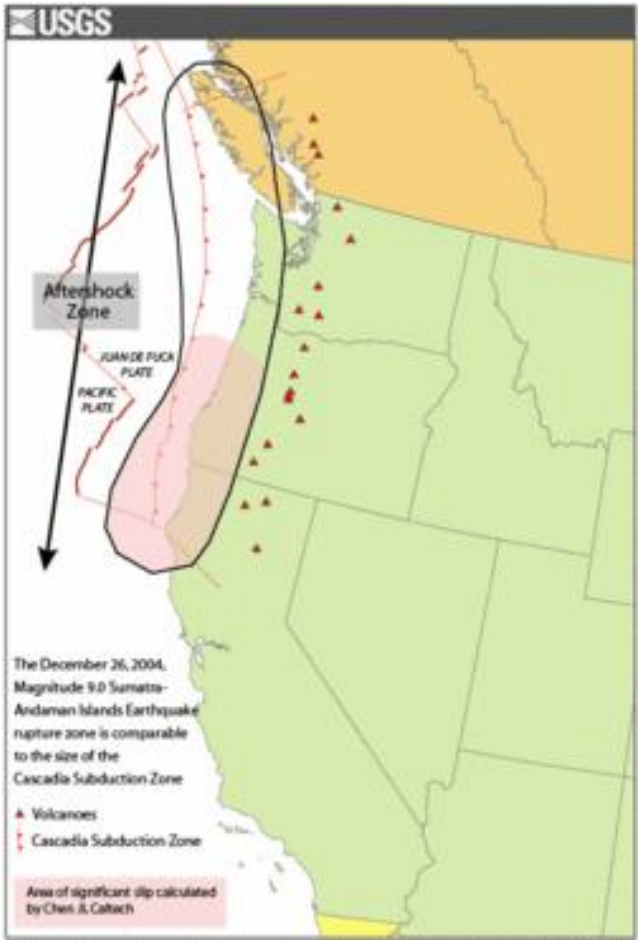
  www.clawpack.org

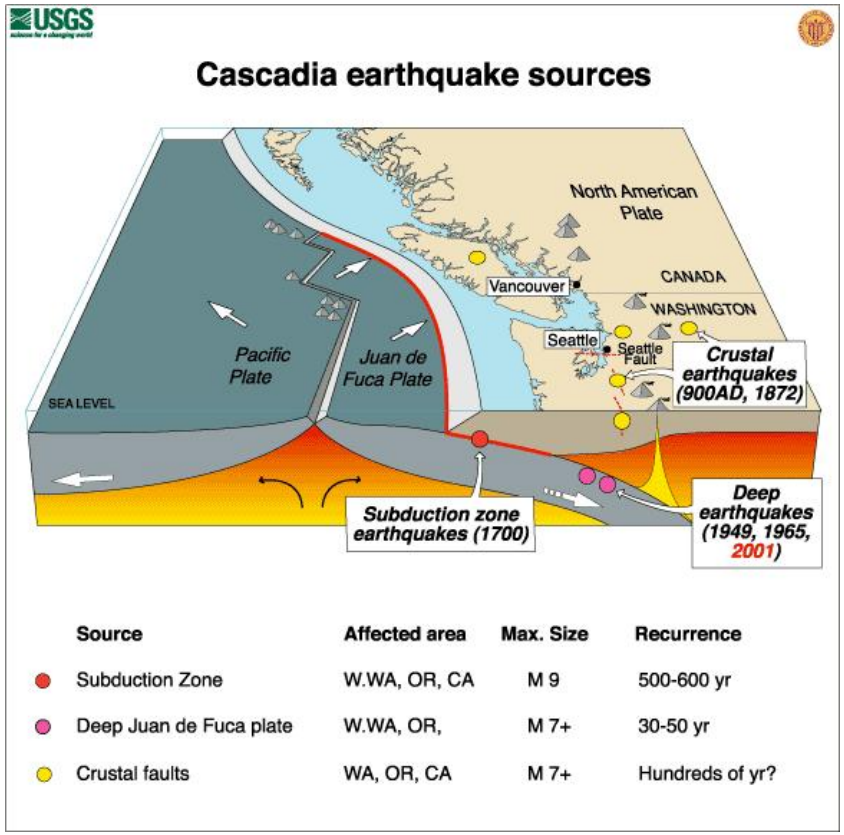
  www.github.com/clawpack

- Since 2003: GeoClaw, geophysical flows and hazards

  Tsunami modeling, storm surge, debris flows, etc.
  Benchmarking codes,
  Hazard assessment projects for Washington State,
  Probabilistic Tsunami Hazard Assessment (PTHA)

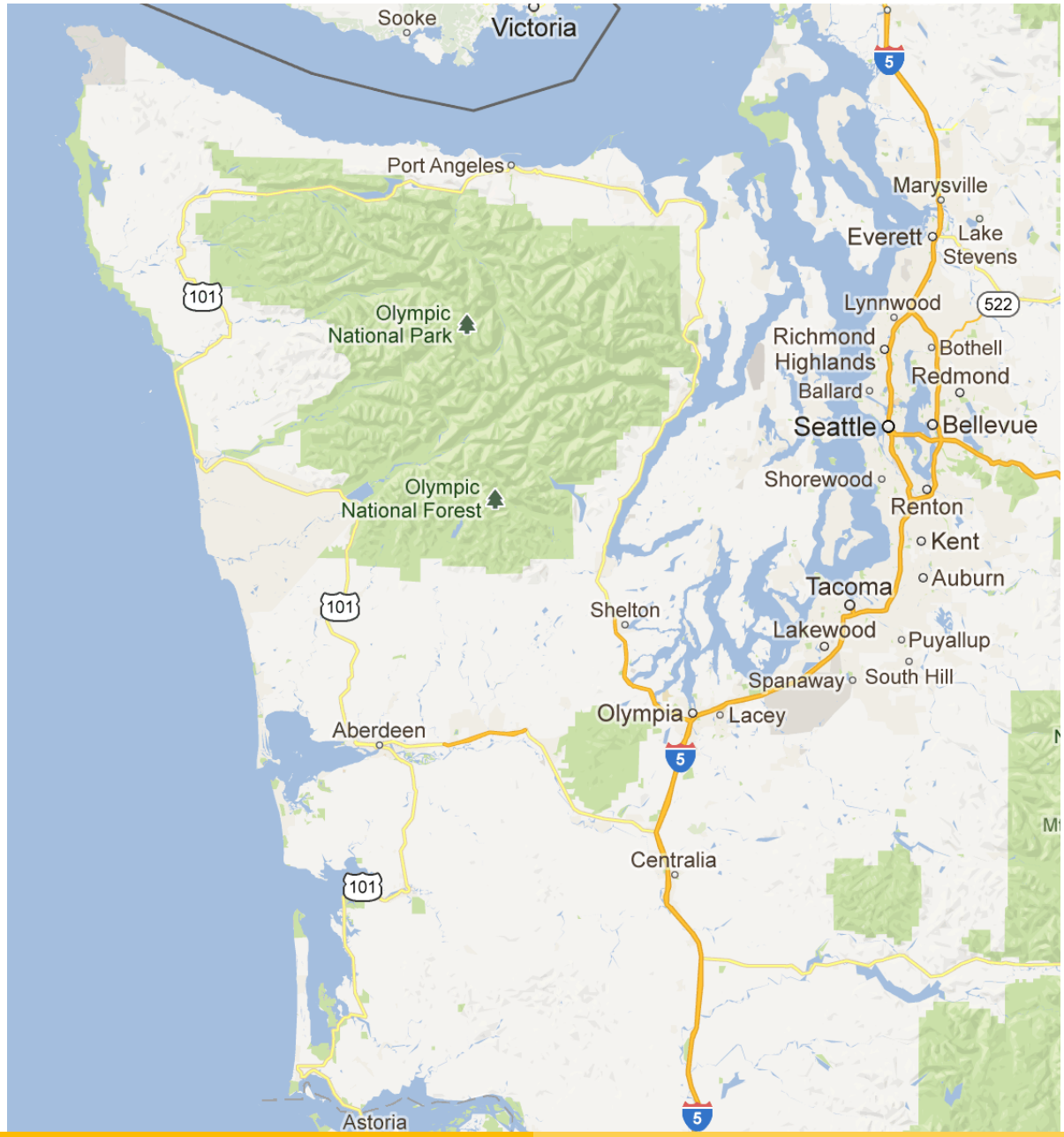
Cascadia Subduction Zone (CSZ)



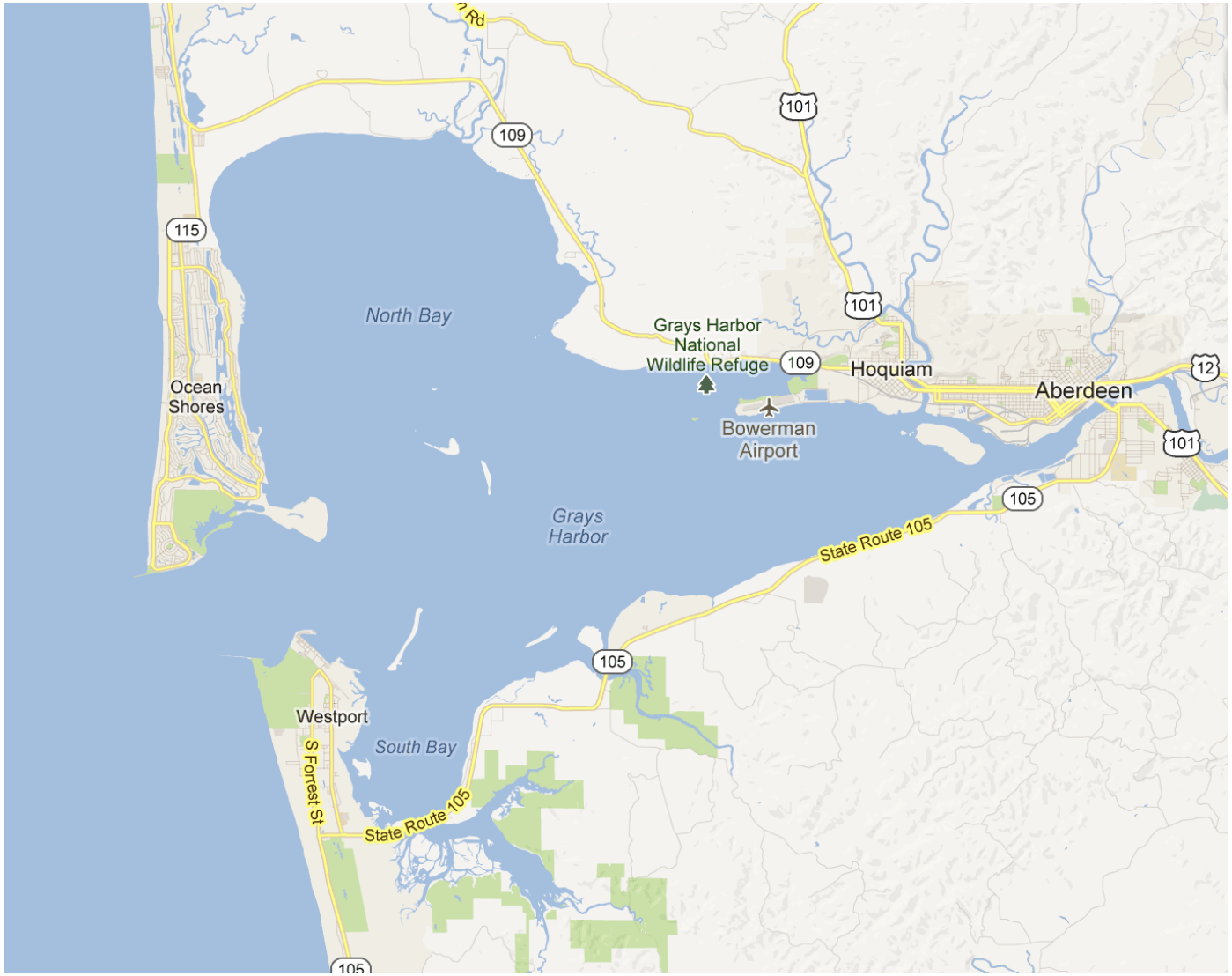


Magnitude 9 event in 1700, mean inter-event time ≈ 500 years

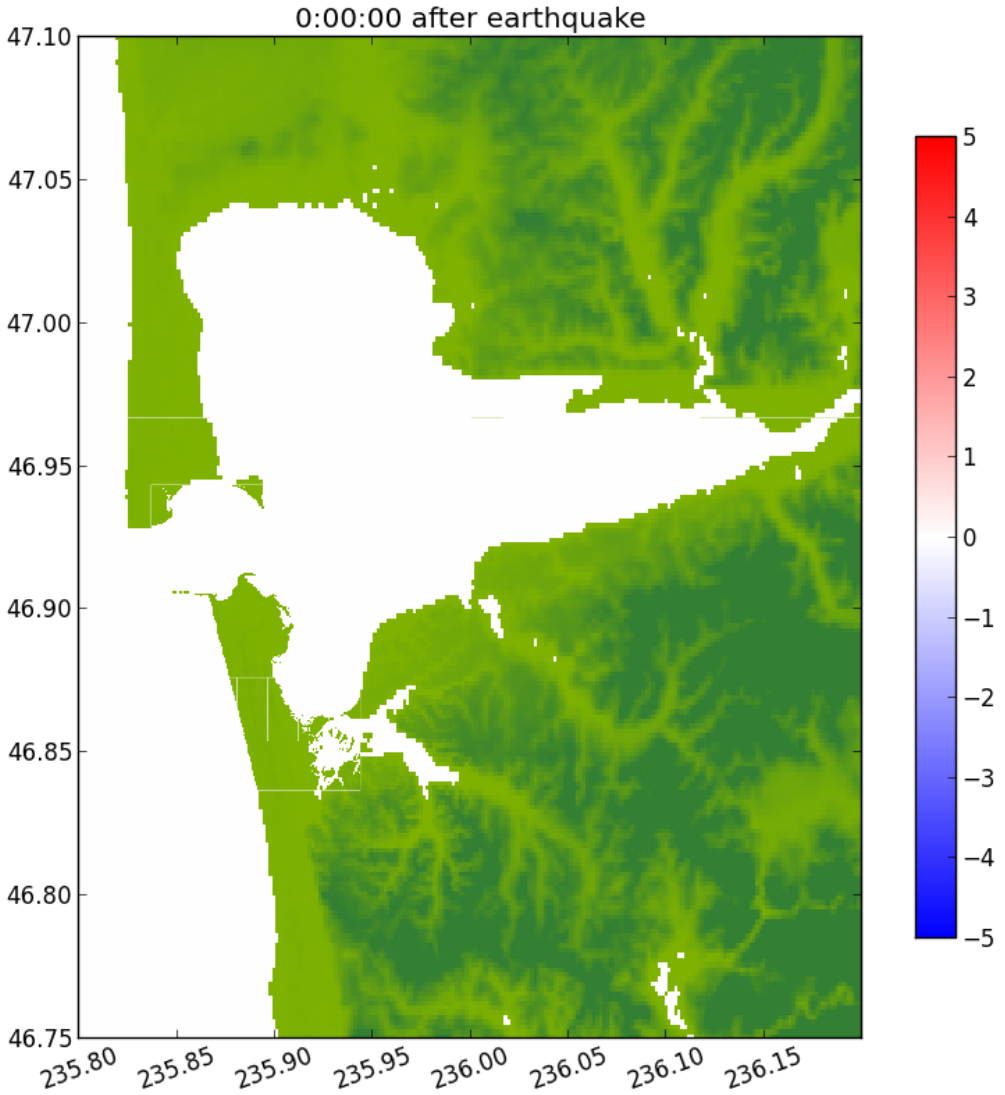
Washington Coast



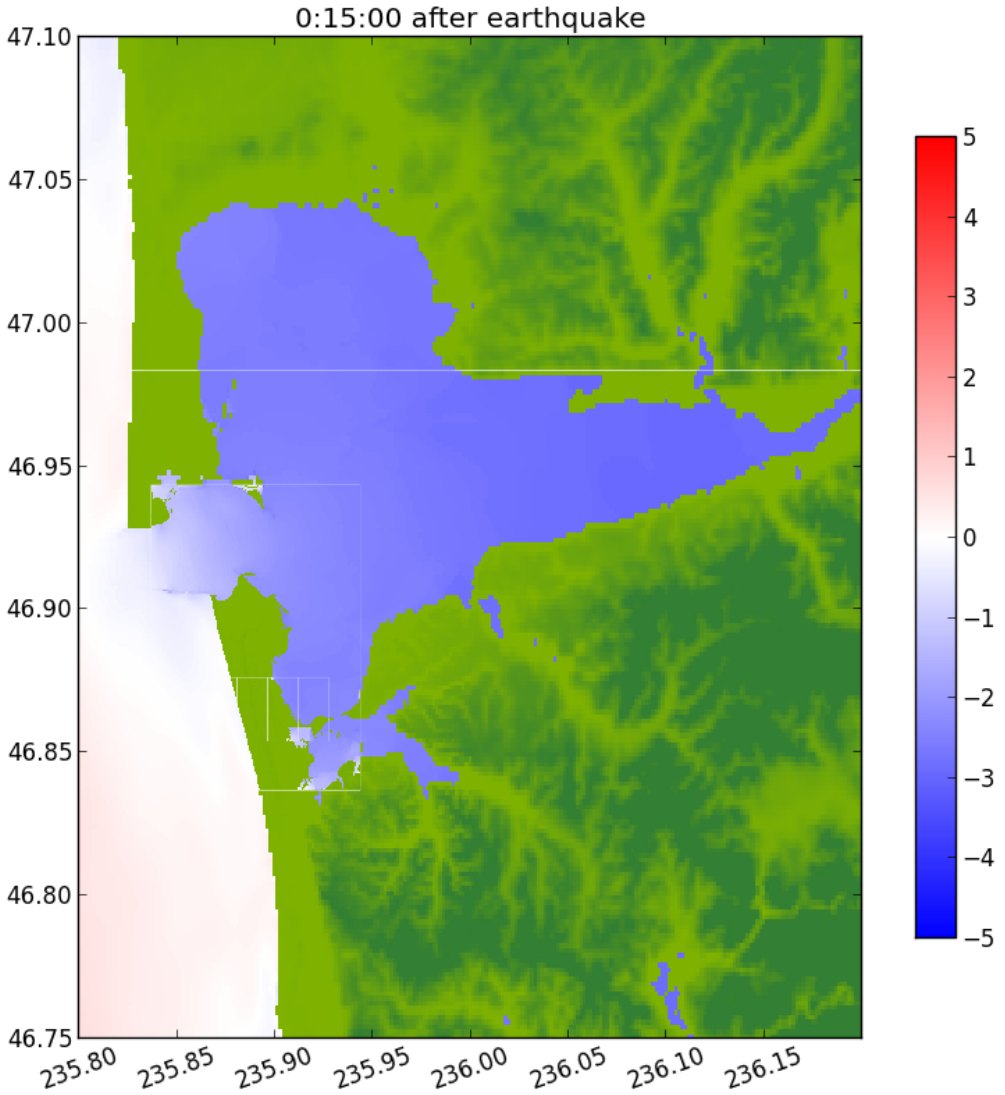
Grays Harbor



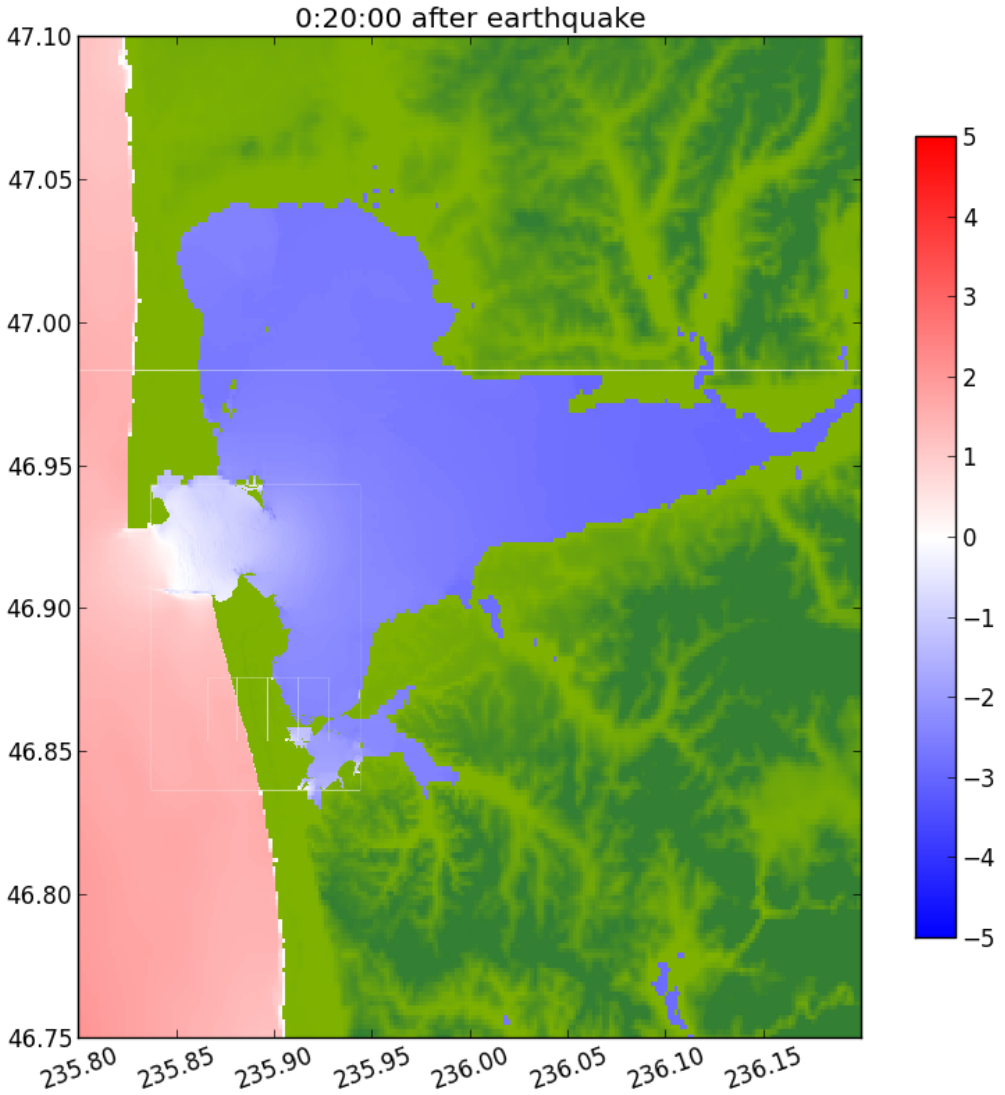
Mw 9.0 Cascadia event hitting Gray's Harbor



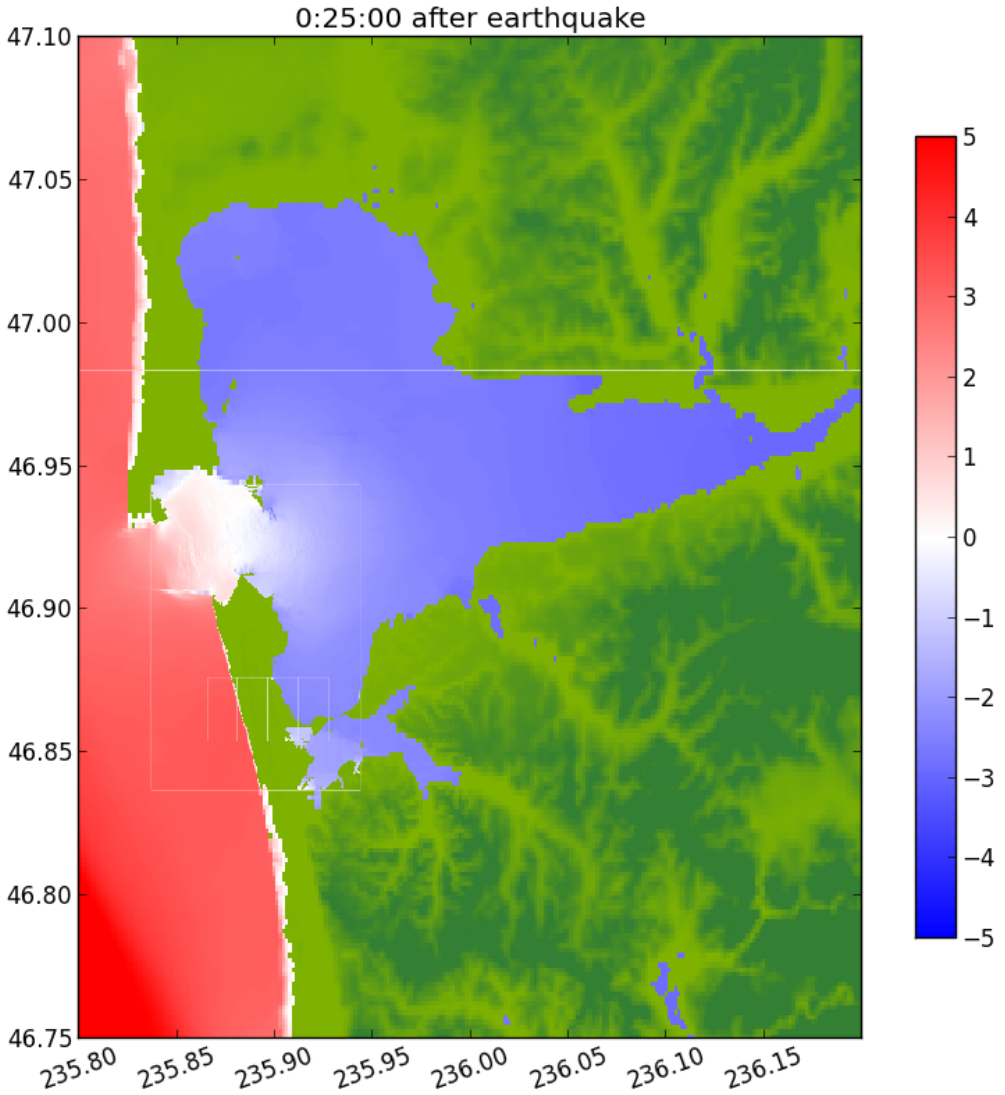
Mw 9.0 Cascadia event hitting Gray's Harbor



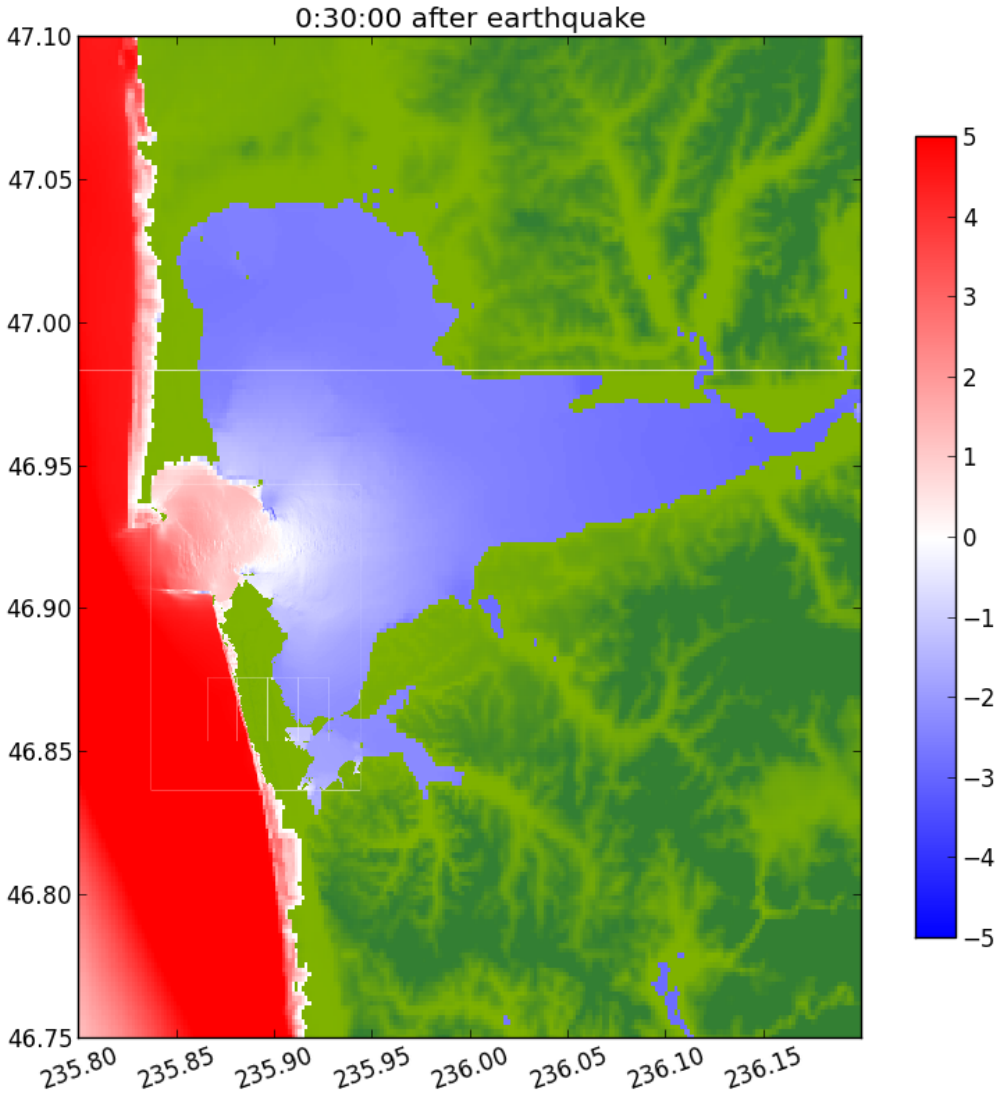
Mw 9.0 Cascadia event hitting Gray's Harbor



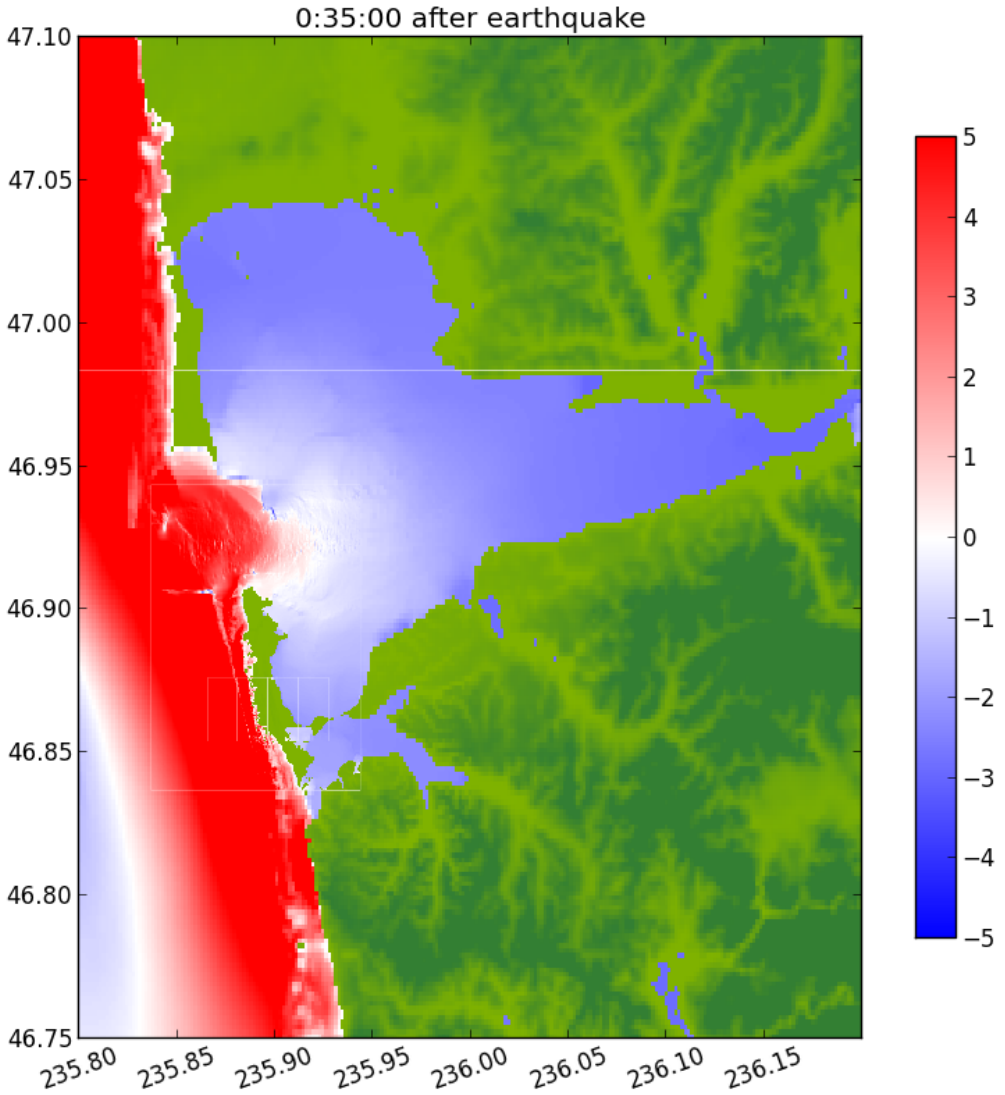
Mw 9.0 Cascadia event hitting Gray's Harbor



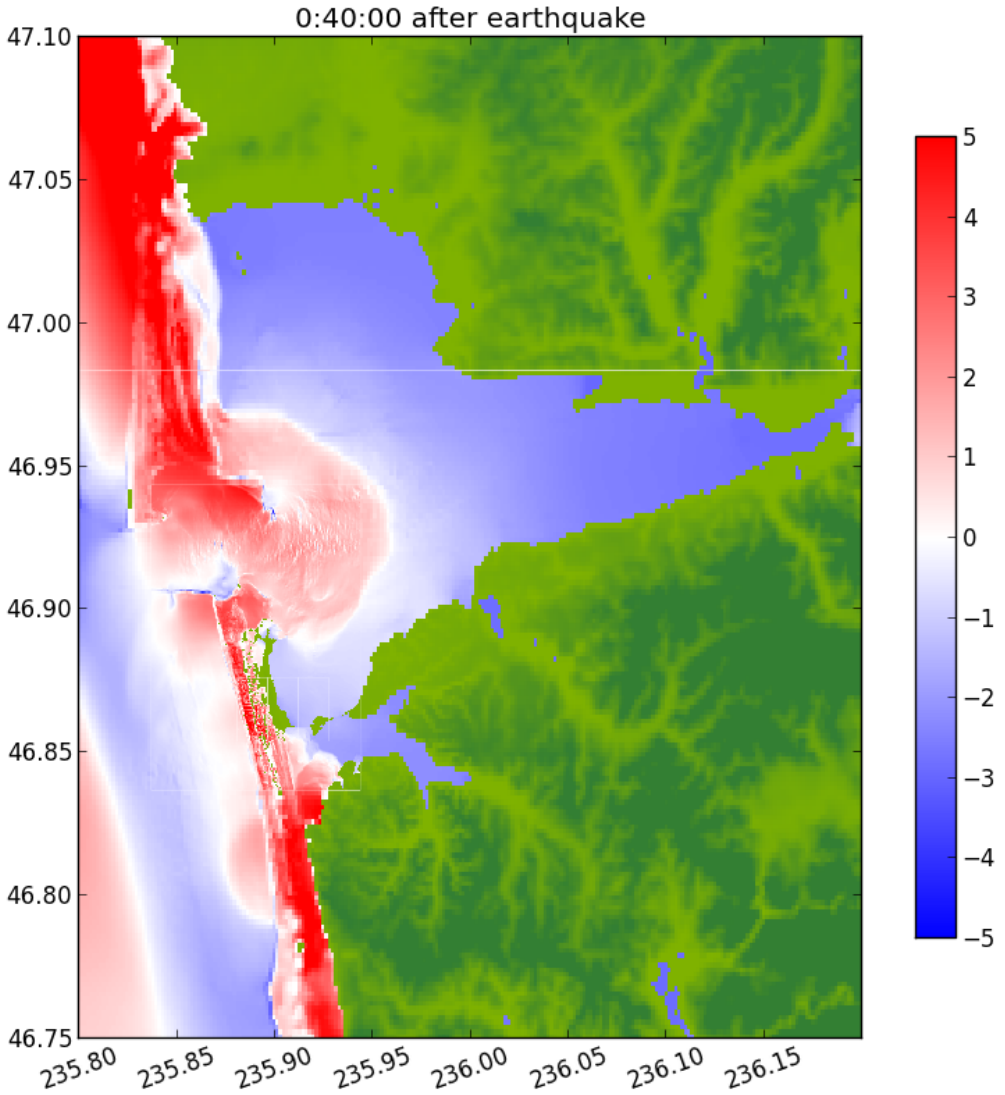
Mw 9.0 Cascadia event hitting Gray's Harbor



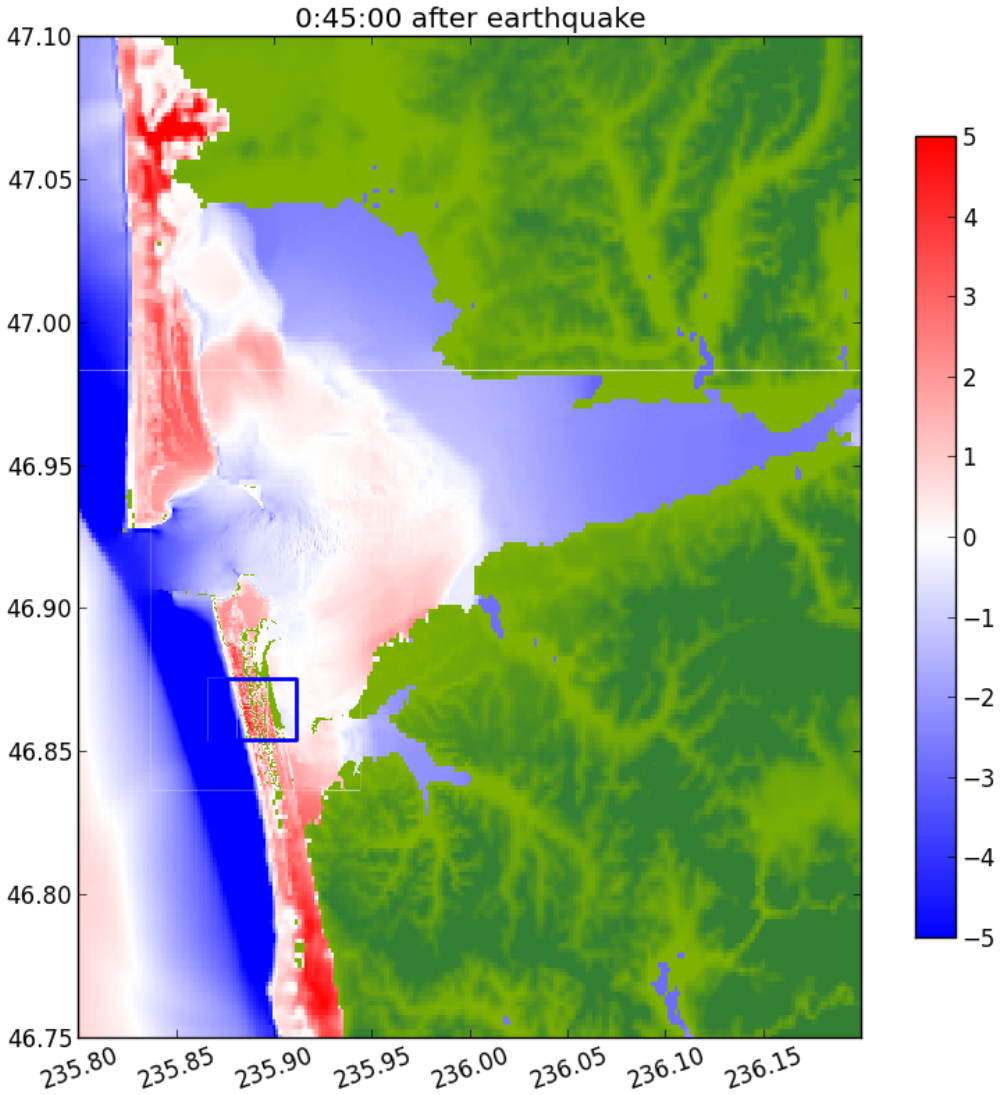
Mw 9.0 Cascadia event hitting Gray's Harbor

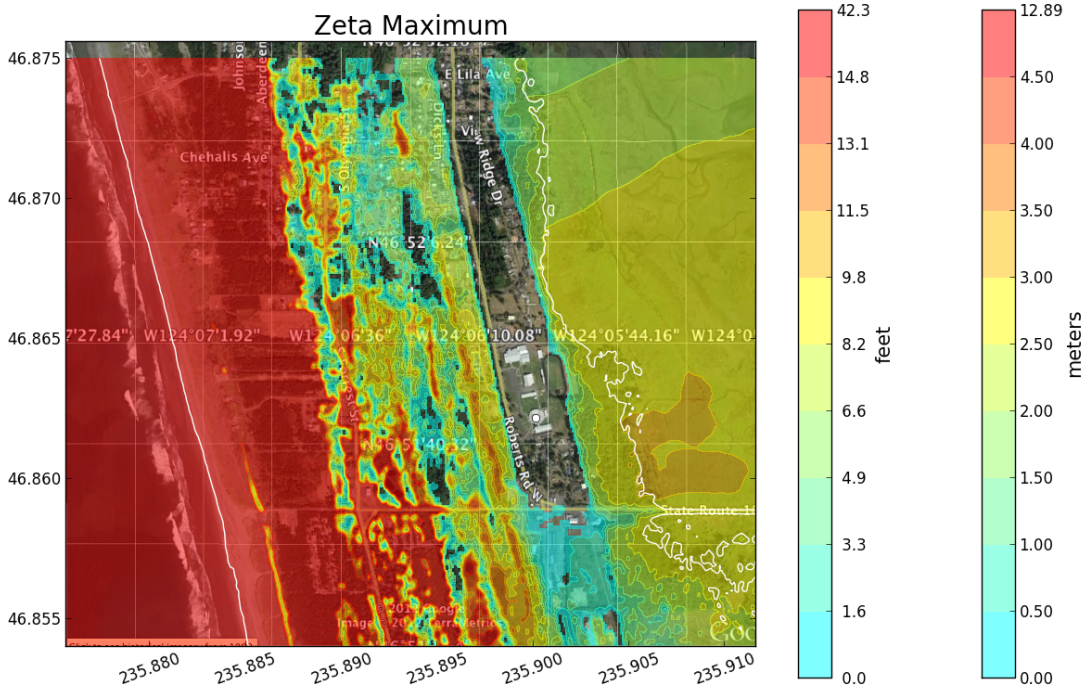


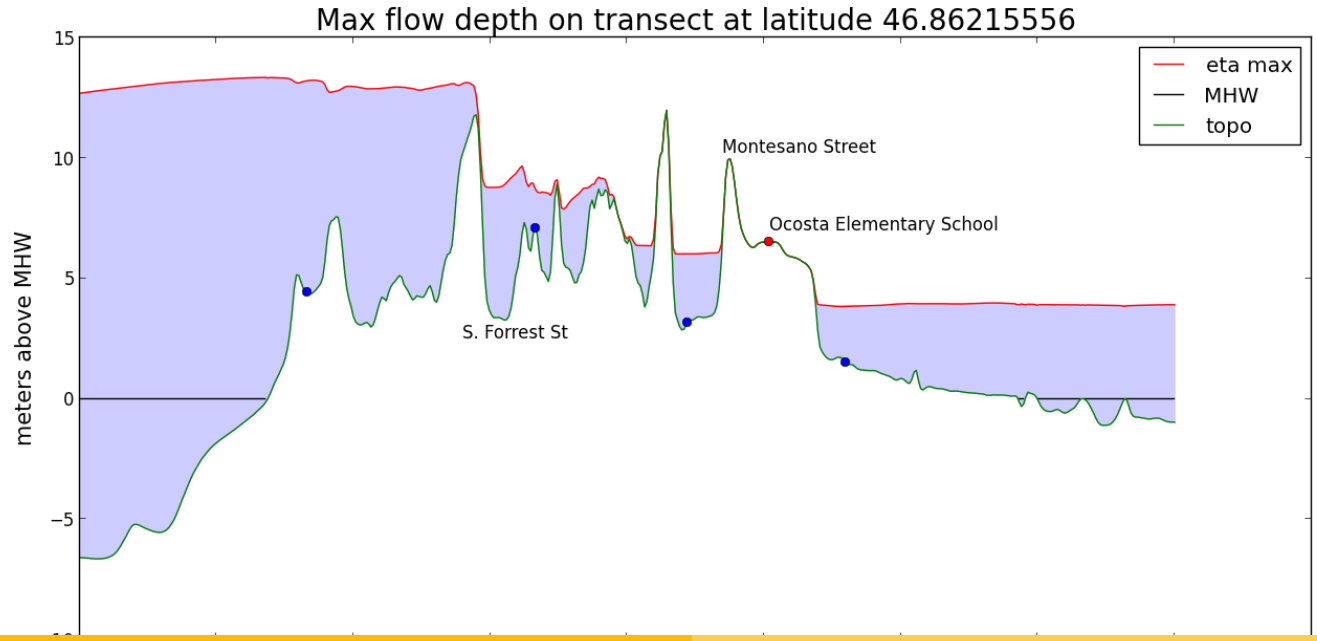
Mw 9.0 Cascadia event hitting Gray's Harbor



Mw 9.0 Cascadia event hitting Gray's Harbor

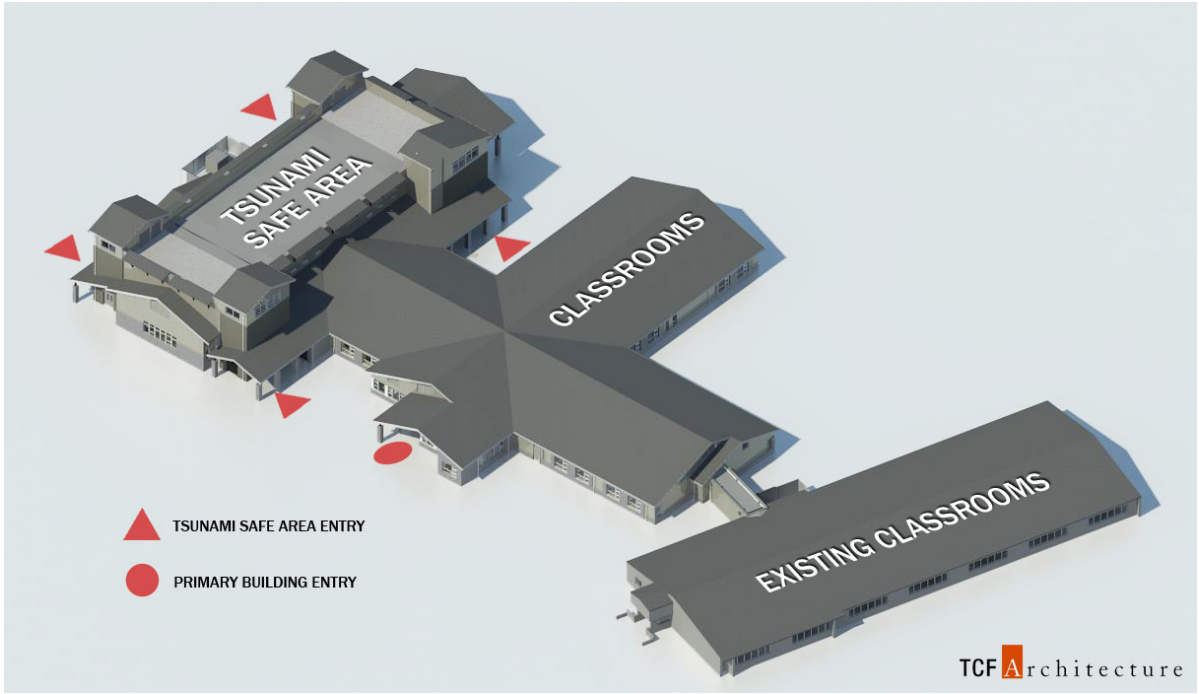



Zeta Maximum
46.875
46.870
46.865
46.860
46.855
235.880
235.885
235.890
235.895
235.900
235.905
235.910
Chehalis Ave
E Lila Ave
View Ridge Dr
Roberts Rd W
W124°07'1.92"
W124°06'10.08"
W124°05'44.16"
42.3
14.8
13.1
11.5
9.8
8.2
6.6
4.9
3.3
1.6
0.0
feet
12.89
4.50
4.00
3.50
3.00
2.50
2.00
1.50
1.00
0.50
0.00
meters


Max flow depth on transect at latitude 46.86215556
meters above MHW
15
10
5
0
−5
eta max
MHW
topo
Montesano Street
Ocosta Elementary School
S. Forrest St

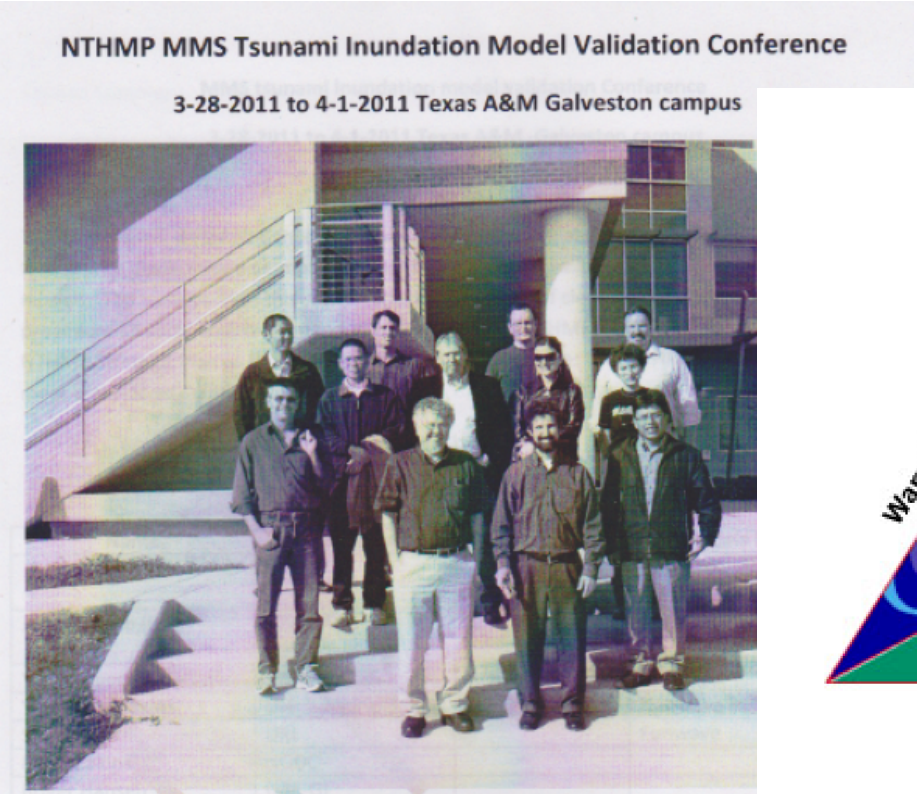
First vertical evacuation structure in US

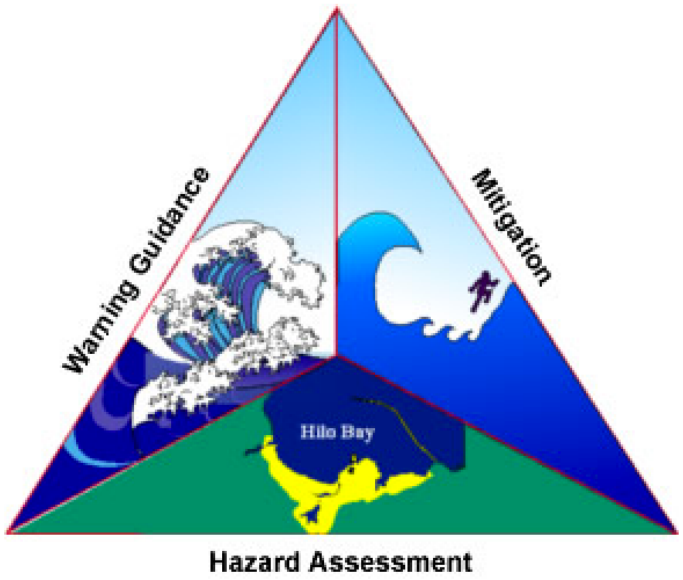
Ocosta Elementary School, Westport, WA



Designed by TCF Architecture of Tacoma, WA, with structural engineering work by Degenkolb Engineers.

National Tsunami Hazard Mitigation Program





Benchmark data can now be found at
github.com/rjleveque/nthmp-benchmark-problems

Guidelines for Reproducibility & Open Science

See: **http://uwescience.github.io/reproducible**

Goals:

- To achieve greater scientific validity and integrity by making it easier to verify published results.
- To increase productivity of current and future researchers on funded projects.
- To increase the impact of the research performed, software developed, and papers published.
- To help promote data and code as first class research products.
- To increase access to and usability of research products by other researchers.
- To use the DSE as a test bed for developing and promoting tools and cultural changes across a broad spectrum of academic disciplines.

What does Reproducible Research mean?

Ability to determine exactly how scientific results were obtained.

- Basis of scientific method.
- Required for confidently building on past results.
- Critical for accountability in engineering analysis / decision making.

What does Reproducible Research mean?

Ability to determine exactly how scientific results were obtained.

- Basis of scientific method.
- Required for confidently building on past results.
- Critical for accountability in engineering analysis / decision making.

Standards and best practices in computational/data science are not yet well codified.

What does Reproducible Research mean?

Ability to determine exactly how scientific results were obtained.

- Basis of scientific method.
- Required for confidently building on past results.
- Critical for accountability in engineering analysis / decision making.

Standards and best practices in computational/data science are not yet well codified.

- Experimental science: Lab notebooks, methodology section of publications, etc.

What does Reproducible Research mean?

Ability to determine exactly how scientific results were obtained.

- Basis of scientific method.
- Required for confidently building on past results.
- Critical for accountability in engineering analysis / decision making.

Standards and best practices in computational/data science are not yet well codified.

- Experimental science: Lab notebooks, methodology section of publications, etc.
- Mathematics: Proofs are required in publications.

Quote from Reproducible Research: A Cautionary Tale

By David Crotty, March 26, 2014 on **the scholarly kitchen blog**

If your experiment consists of running numerical data through an algorithm, then releasing your data and your code allows others to quickly verify that you've done what you've said you've done. But when it comes to other types of research, wet bench experiments or observational work for example, reproduction is not quite so simple.

Quote from Reproducible Research: A Cautionary Tale

By David Crotty, March 26, 2014 on **the scholarly kitchen blog**

If your experiment consists of running numerical data through an algorithm, then releasing your data and your code allows others to quickly verify that you've done what you've said you've done. But when it comes to other types of research, wet bench experiments or observational work for example, reproduction is not quite so simple.

If only it were so easy in computational/data science!

http://www.phdcomics.com/comics/archive.php?comicid=1531

Private reproducibility...

Use scripts, not GUIs, for data analysis and visualization.

Use version control / provenance tracking tools.

Archive code and data used for published results.

Private reproducibility...

Use scripts, not GUIs, for data analysis and visualization.

Use version control / provenance tracking tools.

Archive code and data used for published results.

Why?

- Ability to check results in prior publication,
- Ability to build on your own past research of your own (or students / collaborators).
- Easily modify tables/figures to satisfy referees, etc.

Private reproducibility...

Use scripts, not GUIs, for data analysis and visualization.

Use version control / provenance tracking tools.

Archive code and data used for published results.

Why?

- Ability to check results in prior publication,
- Ability to build on your own past research of your own (or students / collaborators).
- Easily modify tables/figures to satisfy referees, etc.

Auditable Research: Even if code and data are not shared, there should be a permanent record that can be checked.

Analogous to lab notebooks.

Public Reproducibility...

Allowing others to reproduce your results.

(Readers, referees, researchers down the hall...)

Public Reproducibility...

Allowing others to reproduce your results.
(Readers, referees, researchers down the hall...)

Why?

- Verifying scientific integrity of results.
- Aids in understanding ideas, implementing methods
- Increases impact of work.

What does Reproducible Research mean?

What does it mean that others can reproduce your results?

What does Reproducible Research mean?

What does it mean that others can reproduce your results?

Possible answers...

- Download the code and type `make plots`, see identical plots appear.

What does Reproducible Research mean?

What does it mean that others can reproduce your results?

Possible answers...

- Download the code and type `make plots`, see identical plots appear.
- Be able to implement the algorithm from description in paper and other archived sources, and get essentially the same results.

What does Reproducible Research mean?

What does it mean that others can reproduce your results?

Possible answers...

- Download the code and type `make plots`, see identical plots appear.
- Be able to implement the algorithm from description in paper and other archived sources, and get essentially the same results.
- Various things in between.

What does Reproducible Research mean?

What does it mean that others can reproduce your results?

Possible answers...

- Download the code and type `make plots`, see identical plots appear.
- Be able to implement the algorithm from description in paper and other archived sources, and get essentially the same results.
- Various things in between.

Terms such as replicable or repeatable are sometimes used in addition to reproducible.

Tools to facilitate reproducibility

- Version control systems (VCS)

  CVS, Subversion (server-client),
  Git, Mercurial, Bazaar, etc. (distributed).

Tools to facilitate reproducibility

- Version control systems (VCS)

  CVS, Subversion (server-client),
  Git, Mercurial, Bazaar, etc. (distributed).

- Public hosting sites for VCS repositories

  Github, Bitbucket, Google code, sourceforge, etc.

  Collaboration on open source projects,
  Archiving code used for publications.

Tools to facilitate reproducibility

- Version control systems (VCS)

  CVS, Subversion (server-client),
  Git, Mercurial, Bazaar, etc. (distributed).

- Public hosting sites for VCS repositories

  Github, Bitbucket, Google code, sourceforge, etc.

  Collaboration on open source projects,
  Archiving code used for publications.

- Other archives with stable URLs, DOIs

  Institutional or public data repositories,
  journal supplementary materials,
  Figshare, Zenodo, etc.

Tools to facilitate reproducibility

- Workflow Management Systems

  VisTrails, Madagascar, Sumatra, Taverna, Galaxy, etc.

  Capture the workflow used to generate figures, tables, etc.

  Facilitate tracking the provinance of individual results.
  Data, code, compilers, graphics tools, etc.

  Often work together with VCS for source code.

Tools to facilitate reproducibility

- Literate Programming tools

  CWEB, Doxygen, Sphinx, Sweave, etc.

Tools to facilitate reproducibility

- Literate Programming tools

  CWEB, Doxygen, Sphinx, Sweave, etc.

- Notebooks / Publishing tools

  Mathematica, Maple, Matlab,

  Sage, IPython, knitr, RStudio, etc.

Tools to facilitate reproducibility

- Virtualization

  Package code along with complete environment
  (OS, compilers, graphics tools, etc.)

  E.g., VirtualBox, VMware, etc.

Tools to facilitate reproducibility

- Virtualization

  Package code along with complete environment
  (OS, compilers, graphics tools, etc.)

  E.g., VirtualBox, VMware, etc.

- Cloud computing

  E.g., Amazon EC2, Windows Azure, etc. + VM

Tools to facilitate reproducibility

- Virtualization

  Package code along with complete environment
  (OS, compilers, graphics tools, etc.)

  E.g., VirtualBox, VMware, etc.

- Cloud computing

  E.g., Amazon EC2, Windows Azure, etc. + VM

- Web platforms for running code

  E.g., RunMyCode.org, wakari.io,

  cloud.sagemath.com

Some Git references

- `http://gitref.org/index.html`
- `http://git-scm.com/book/en/Getting-Started-Git-Basics`
- `http://help.github.com/`
- `https://confluence.atlassian.com/display/BITBUCKET/Bitbucket+101`

- **List of 10 recommended tutorials**
- **Github online tutorial**
- **My Coursera class**

Science Code Manifesto

Manifesto Discussion Endorse Resources About

Software is a cornerstone of science. Without software, twenty-first century science would be impossible. Without better software, science cannot progress.

But the culture and institutions of science have not yet adjusted to this reality. We need to reform them to address this challenge, by adopting these five principles:

Code	All source code written specifically to process data for a published paper must be available to the reviewers and readers of the paper.
Copyright	The copyright ownership and license of any released source code must be clearly stated.
Citation	Researchers who use or adapt science source code in their research must credit the code's creators in resulting publications.
Credit	Software contributions must be included in systems of scientific assessment, credit, and recognition.
Curation	Source code must remain available, linked to related materials, for the useful lifetime of the publication.

Endorsements at: `http://sciencecodemanifesto.org/`

Guidelines for Reproducibility & Open Science

See: **http://uwescience.github.io/reproducible**

Goals:

- To achieve greater scientific validity and integrity by making it easier to verify published results.
- To increase productivity of current and future researchers on funded projects.
- To increase the impact of the research performed, software developed, and papers published.
- To help promote data and code as first class research products.
- To increase access to and usability of research products by other researchers.
- To use the DSE as a test bed for developing and promoting tools and cultural changes across a broad spectrum of academic disciplines.

Some links ...

- **2012 ICERM Workshop on Reproducibility**
  Many links on the **wiki** and in **Final report**
- **2011 UBC Workshop with videos online**
- **reproducibleresearch.net**
- **Tutorial: Workflows for reproducible research in computational neuroscience** by Andrew Davison
- **10 Simple Rules for the Care and Feeding of Scientific Data**
  by A. Goodman, A. Pepe, A. W. Blocker, et al.
- **Best Practices for Scientific Computing**
  by G. Wilson, D. A. Aruliah, C. T. Brown, et al.