SAND2016-7121C

Waveform Correlation Detection: Putting Theory into Practice

Megan Slinkard, Amy Sundermier, Stephen Heck, Chris Young

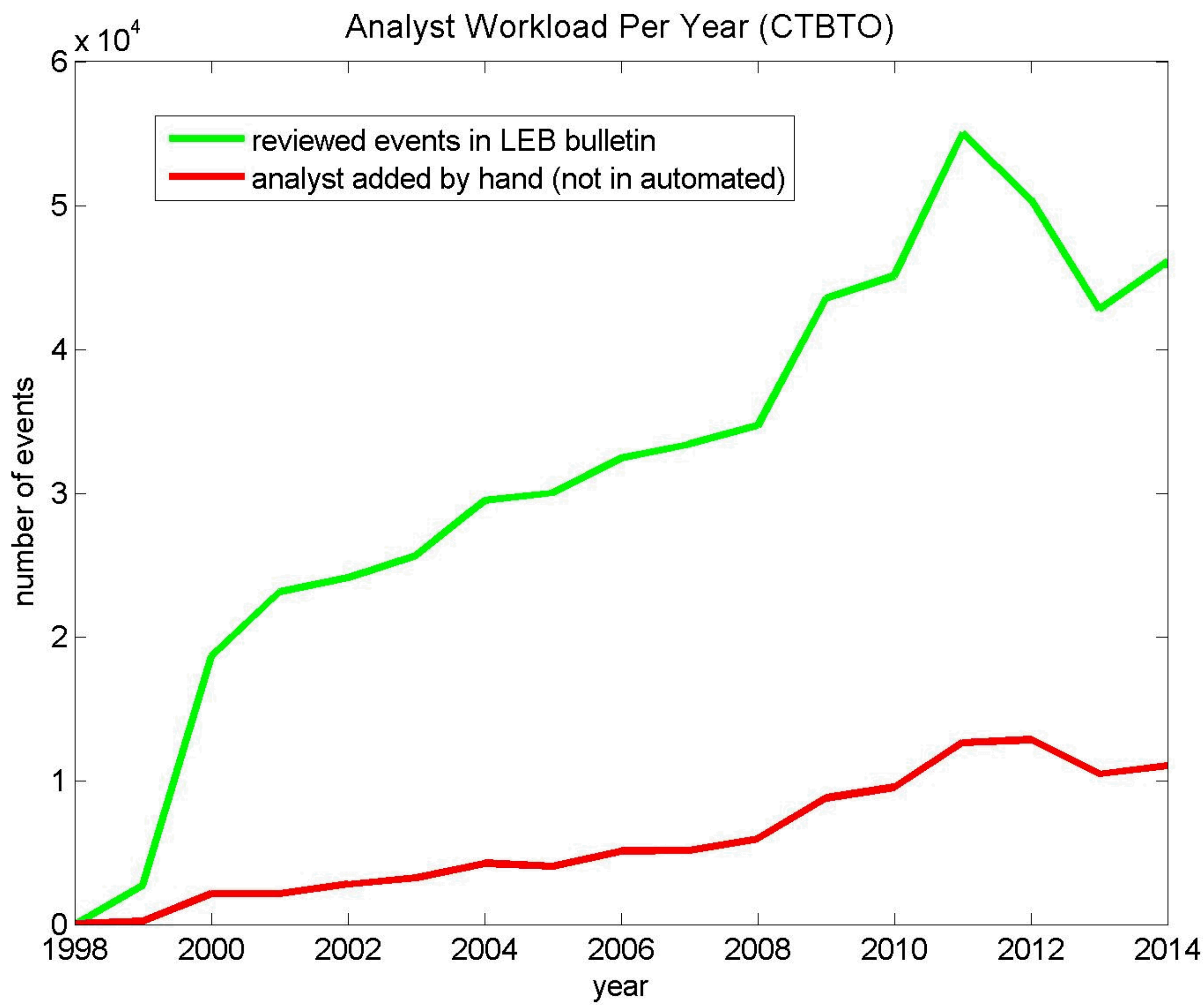
Sandia National Laboratories

Paul Richards, David Schaff

Lamont Doherty Earth Observatory

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Motivation


Analyst Workload Per Year (CTBTO)
x 10^4
6
5
4
3
2
1
0
number of events
1998
2000
2002
2004
2006
2008
2010
2012
2014
year
reviewed events in LEB bulletin
analyst added by hand (not in automated)

Benefits

Benefits of Waveform Correlation

- Only need **1 station**
- Better detection in noise **(~1 Mag unit lower)**
- Initial **location** and **magnitude** values
- More consistent phase picks (leads to more **accurate relative locations**)
- **Real time tracking** of repeating events

- Use these benefits to improve use of **analyst resources**

Objective

Use waveform correlation's advantages in detecting repeated seismicity to improve monitoring efficiency…

While improving detection capability, and location and magnitude estimates.

Questions we focused on

How many events would global waveform correlation detection provide to the monitoring agencies?

- Events in current bulletins, low magnitude events

How will large-scale waveform correlation detection research be accomplished?

- Process *continuous* waveform data across a network
- Developed computational infrastructure to operate at research speed

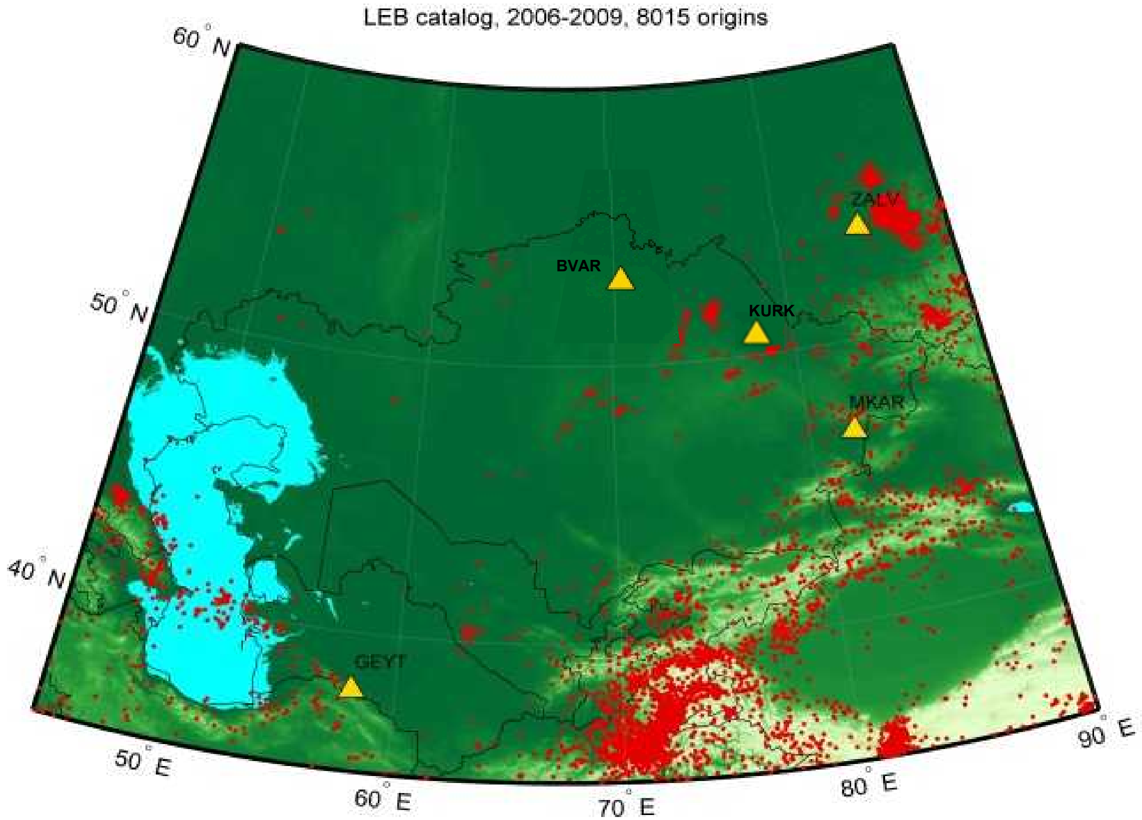
How would an automated system work?

- Need automated processes for selecting templates, setting thresholds, validating results.
- Practical, yet robust.

- **Template approaches – selecting templates, setting thresholds**
- Validating detections
- Our computational infrastructure
- Datasets and results
  - Central Asia – broad-scale monitoring
  - Wenchuan – replicating local performance with regional stations

Scenario #1

- Monitoring a region
- Want templates that span it
- One station

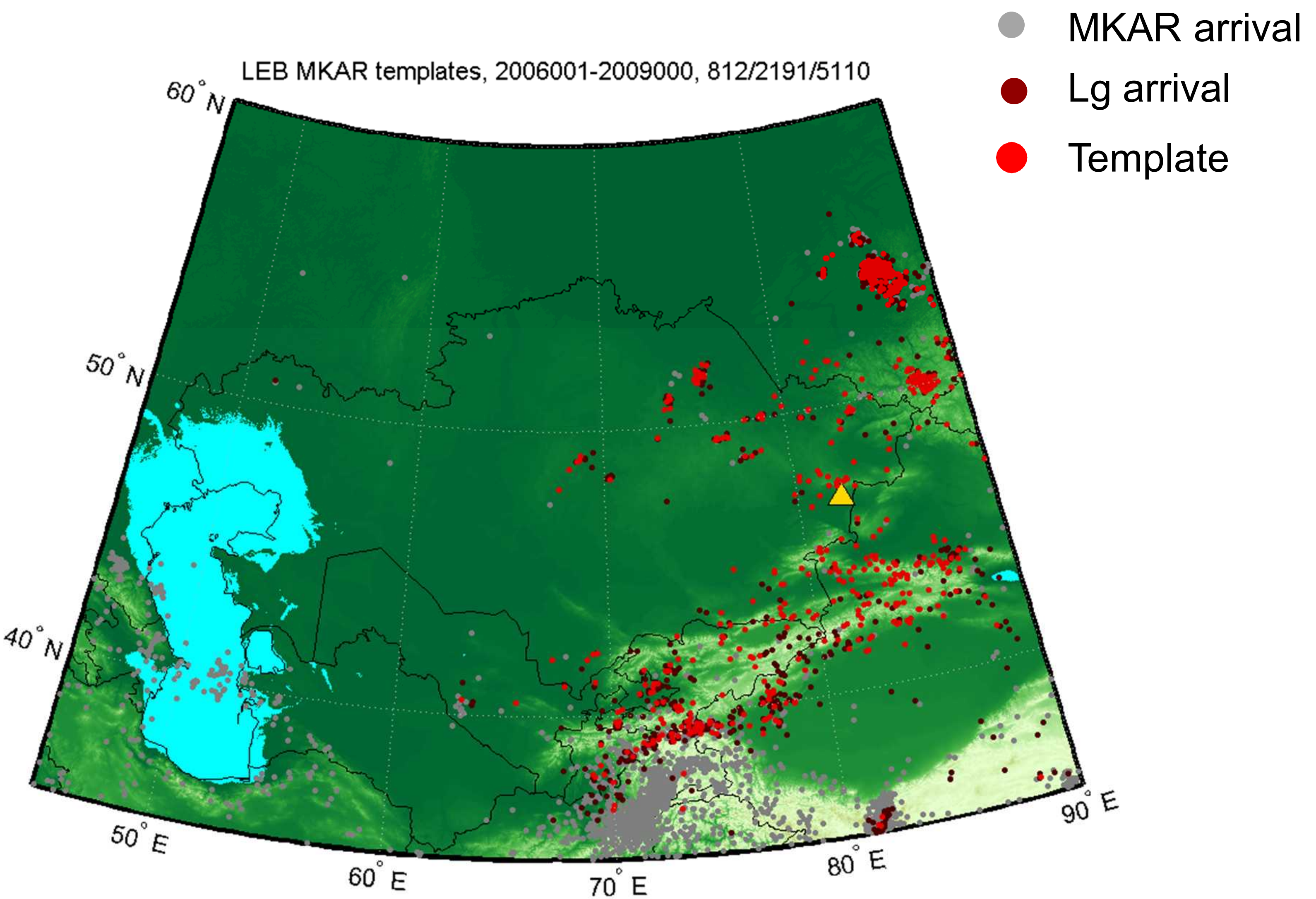


Slinkard, Megan, David Schaff, Natalya Mikhailova, Stephen Heck, Christopher Young, and Paul G. Richards. "Multistation validation of waveform correlation techniques as applied to broad regional monitoring." *Bulletin of the Seismological Society of America* 104, no. 6 (2014): 2768-2781.

Template selection

- Template options:
  - Use historical archive
  - Add events in (near) real time
- For this dataset:
  - Templates from LEB events in 3 year period (2006-2008)
  - Used Lg arrivals
  - Clustered template candidates to get one representative waveform for each mine, etc

Template selection

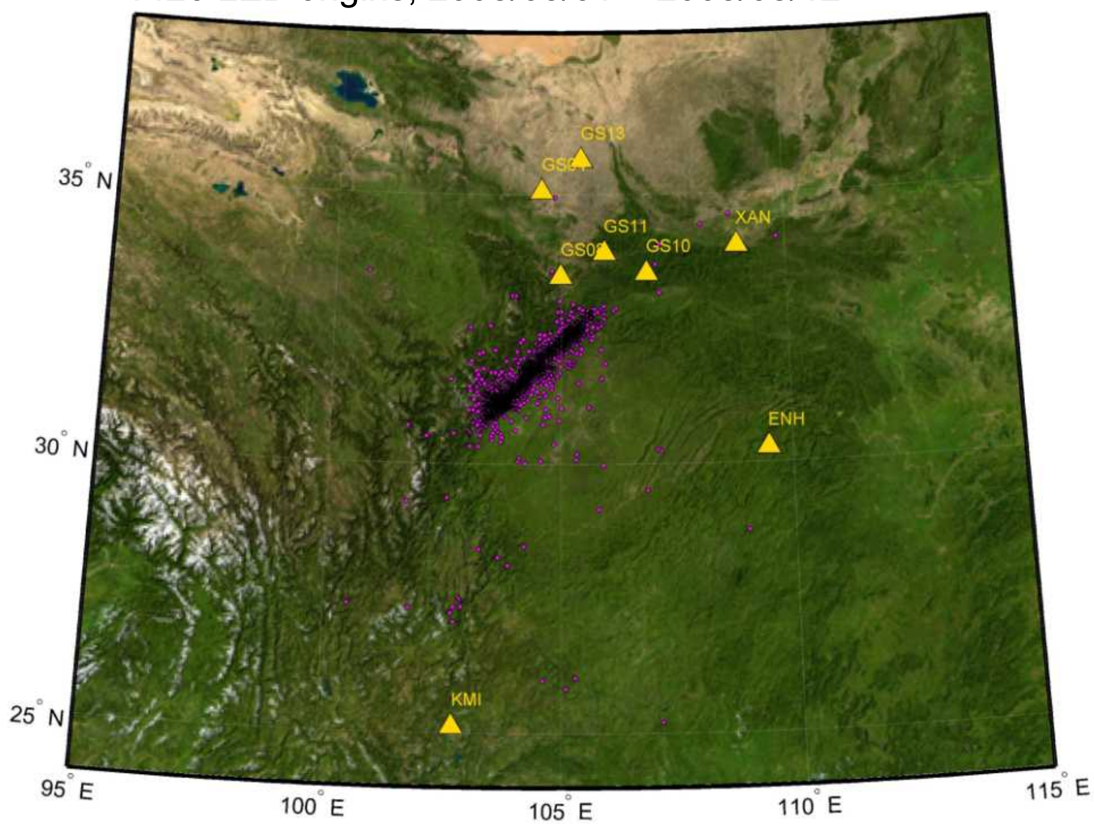


Slinkard, Megan, David Schaff, Natalya Mikhailova, Stephen Heck, Christopher Young, and Paul G. Richards. "Multistation validation of waveform correlation techniques as applied to broad regional monitoring." *Bulletin of the Seismological Society of America* 104, no. 6 (2014): 2768-2781.

Scenario #2

- Focused study of a region
- Stations used are not in IMS network

1426 LEB origins; 2008/05/01 – 2008/08/12



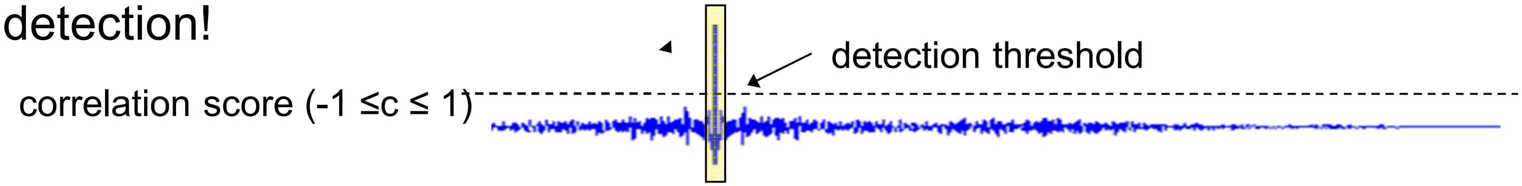
Slinkard, Megan, Stephen Heck, David Schaff, Nedra Bonal, David Daily, Christopher Young, and Paul Richards "Detection of the Wenchuan Aftershock Sequence using Waveform Correlation with a Composite Regional Network." *Bulletin of the Seismological Society of America* (MS# BSSA-D-15-00333, submitted November, 2015)

Template selection

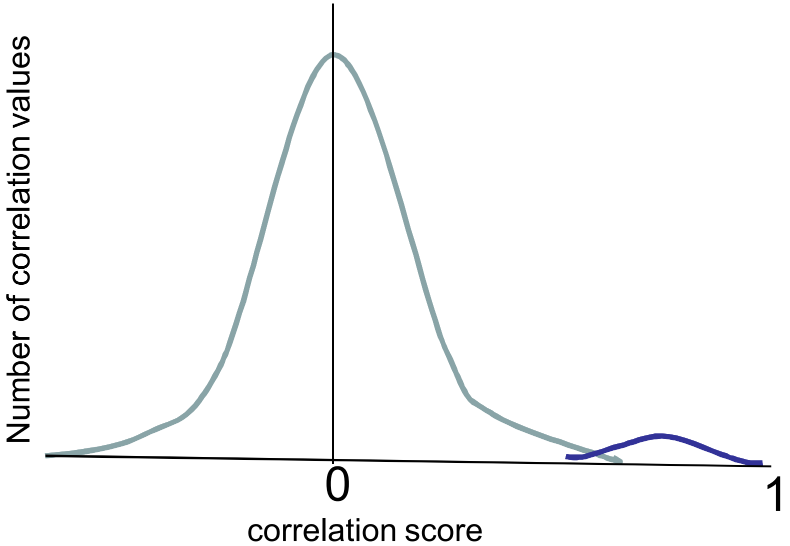
- For this dataset:
  - Templates from all LEB events in 3 month period
  - Because the LEB was not based on any of the stations we processed, we had *no arrival pick information*
  - Used expected Lg arrivals
  - Did no screening or clustering on templates
  - Assumed that each event had a useful template at each station, and, that our threshold setting method would minimize detections from poor templates

Setting Template Correlation Thresholds

- Want a robust, automated, method, based on desired probability of false detection!

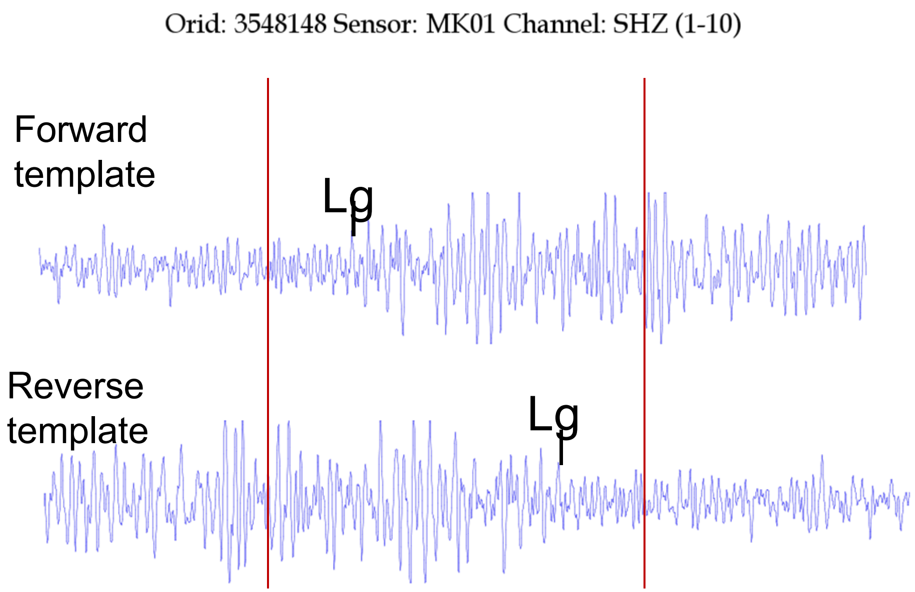


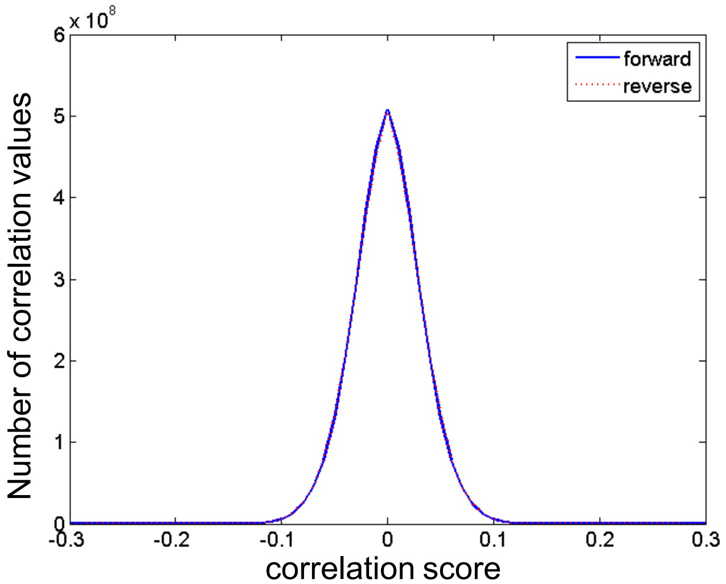
- Distribution of correlation values can be thought of as the sum of 2 distributions: Noise + Events

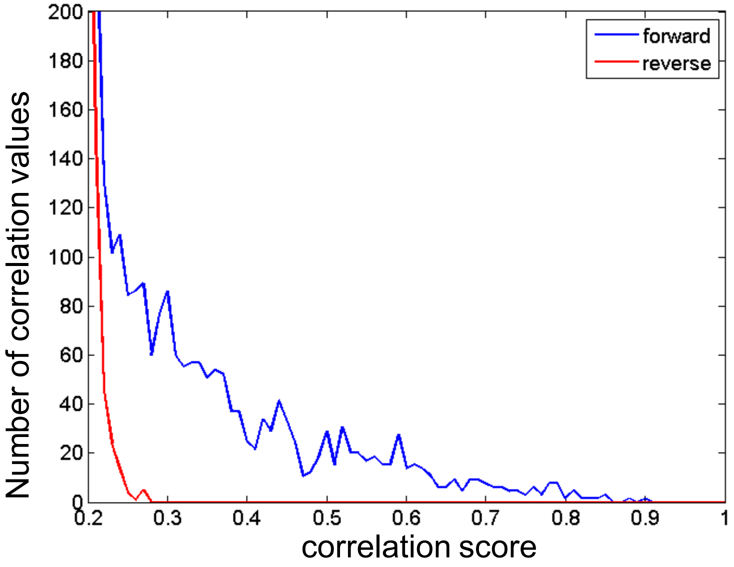


- Knowing distribution of the contribution from noise windows only would allow us to figure out how many false detections to expect at a given correlation threshold.

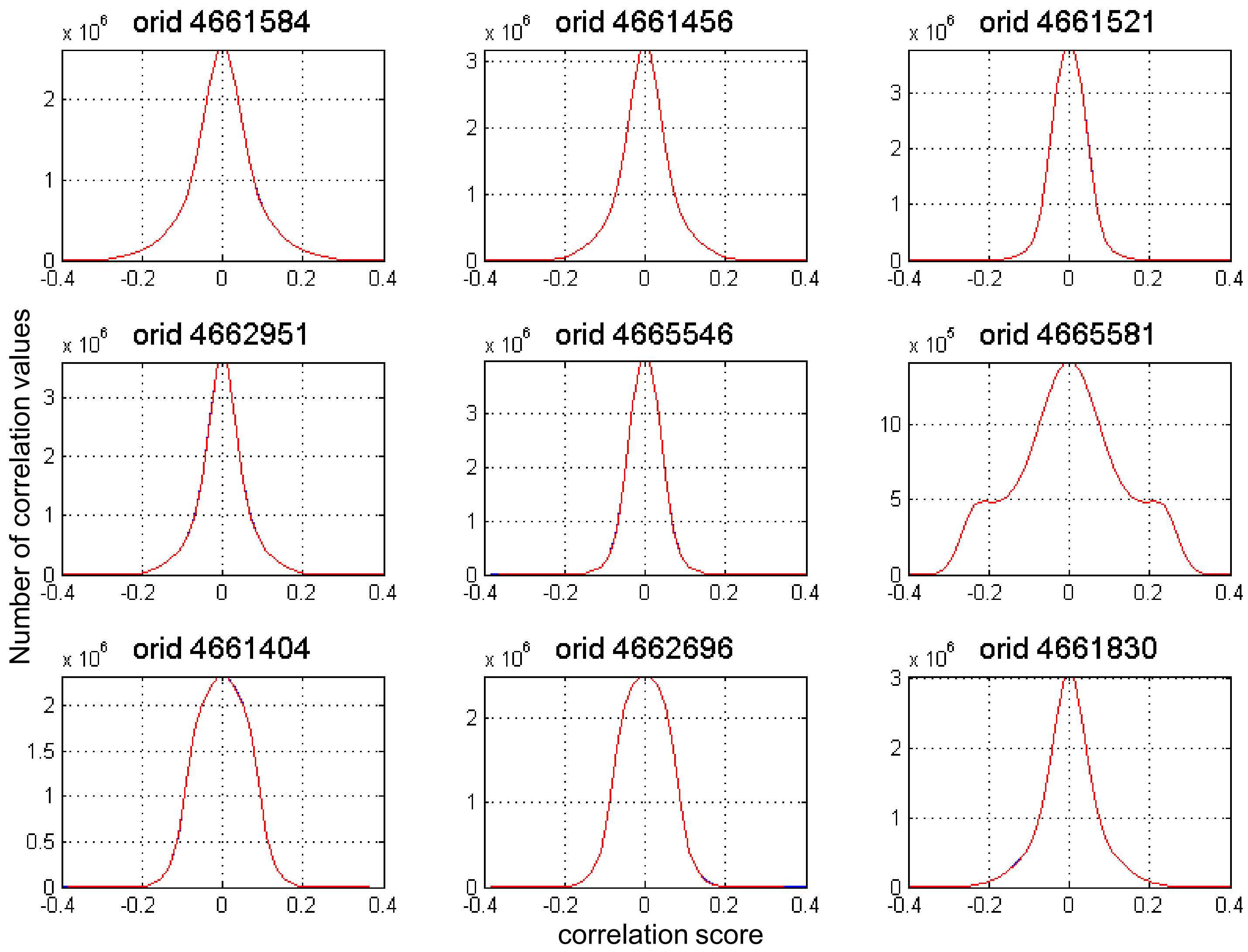
Example


Orid: 3548148 Sensor: MK01 Channel: SHZ (1-10)
Forward template
Lg
Reverse template
Lg

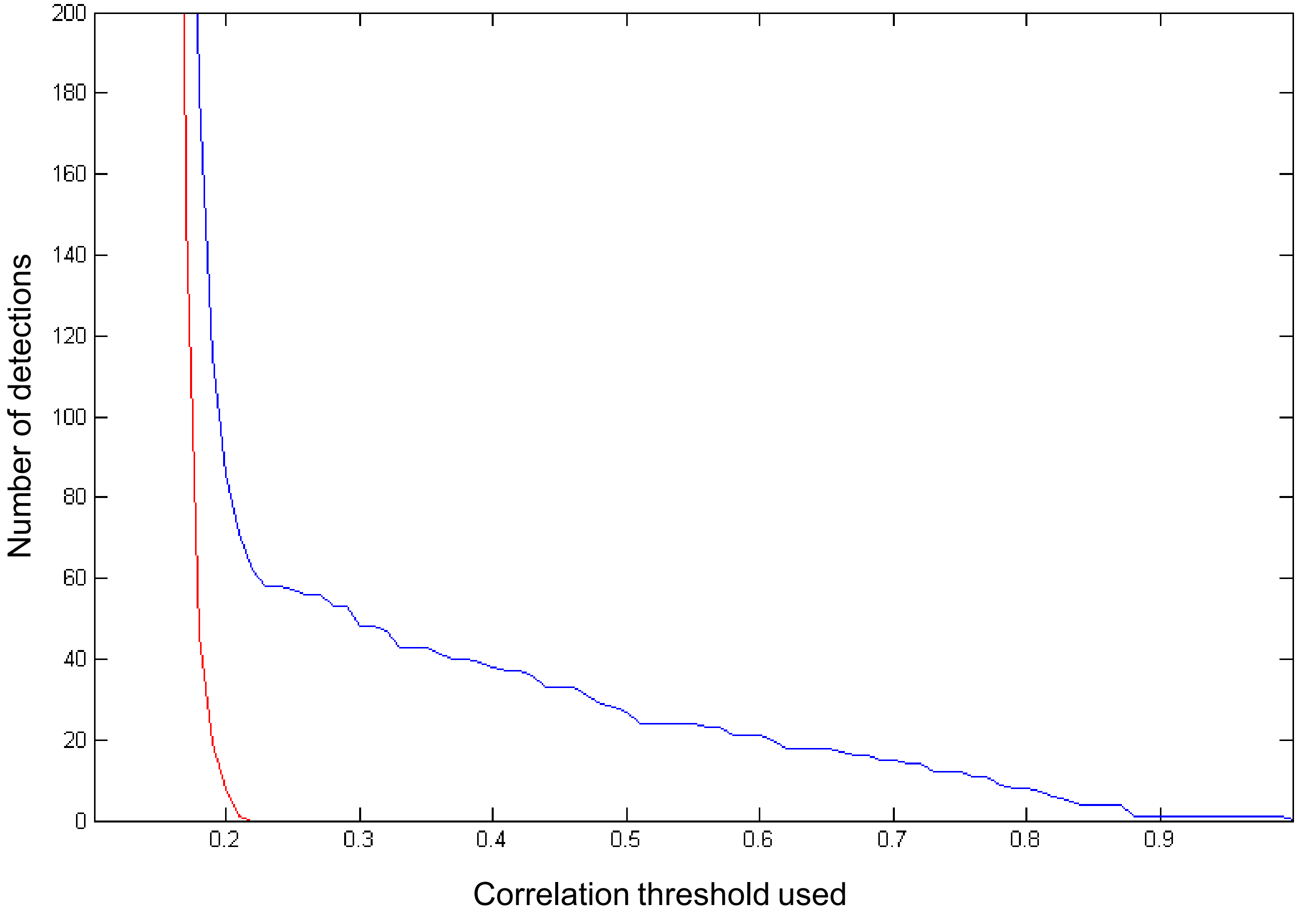

x 10^8
forward
reverse
Number of correlation values
correlation score


forward
reverse
Number of correlation values
correlation score

Examples of Distributions of Correlation Values
from Forward (blue) and Reversed (red) templates



Number of *detections* from Forward and Reverse templates



We can see that the detections from noise start when the correlation threshold has a value of about .22

Using the Time-Reversed Method to set Thresholds

- Use time-reversed template on 3 years of data
- Find the correlation value for which it would have had 3 detections.
- This becomes our threshold, with an expected noise detection rate of 1/yr.

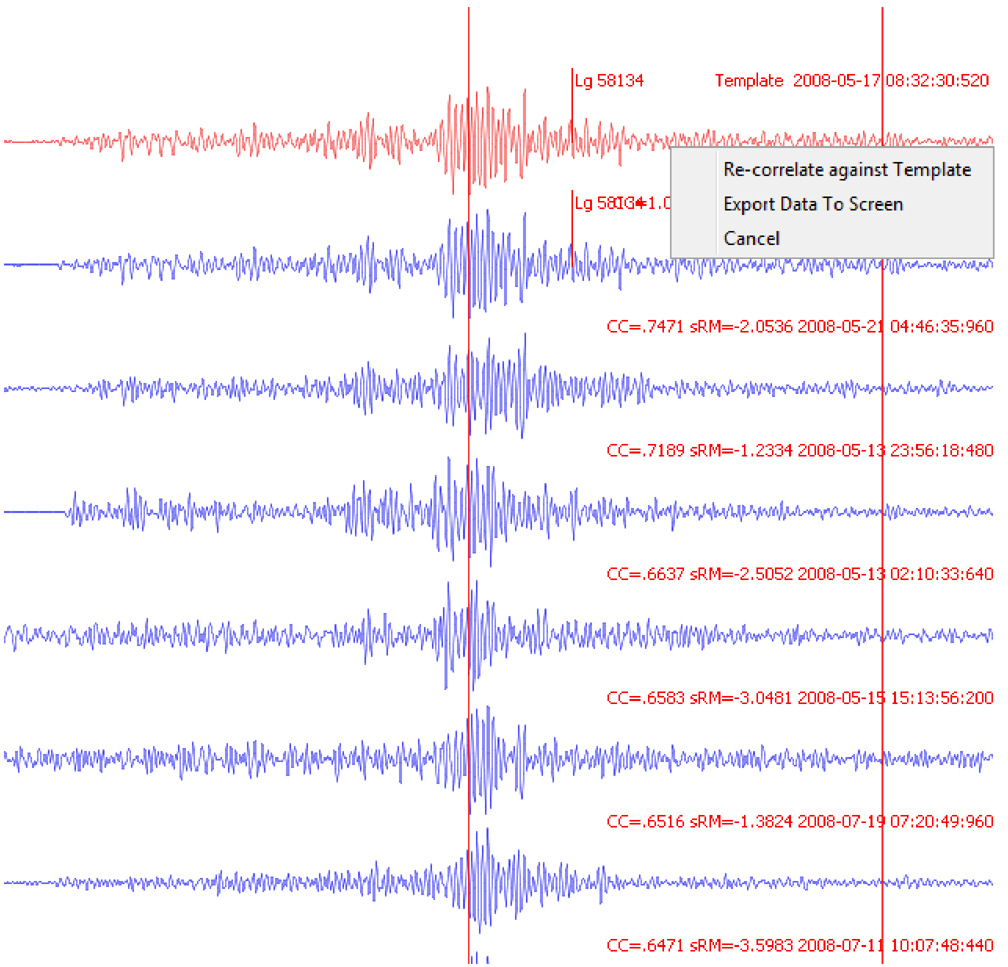
Benefits of setting thresholds this way

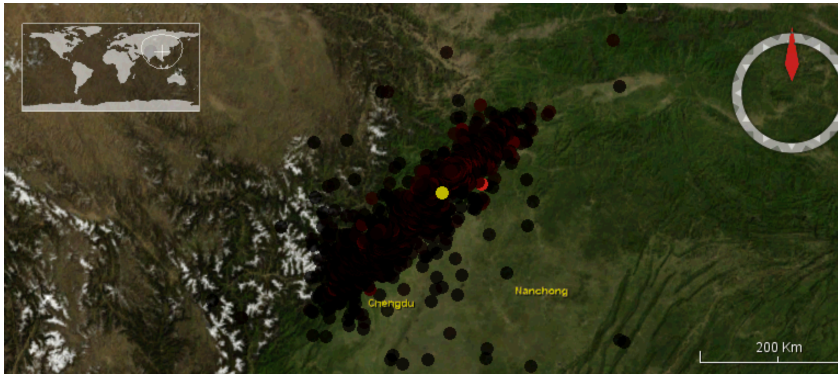
- No assumptions about distribution of noise (Gaussian, etc)
- Can easily make all templates for a station have consistent FAR
- Noisy templates (narrow-band noise, spikes) will be assigned a high threshold -> not a problem

- For templates at MKAR monitoring central Asia, thresholds varied from .16 to .63.
- For templates at GS08 monitoring Wenchuan, thresholds varied from .27 to .58

High threshold (.58)

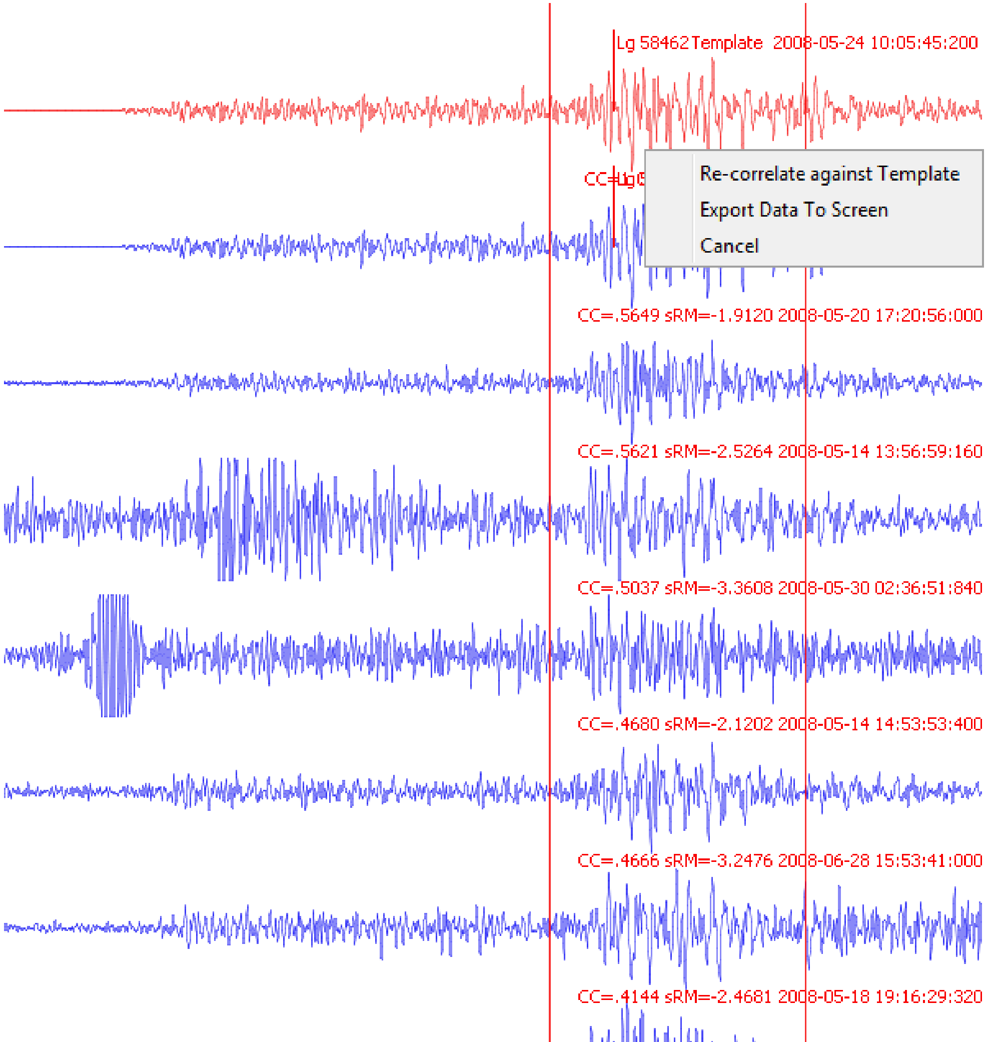
- Window not centered on Lg arrival…

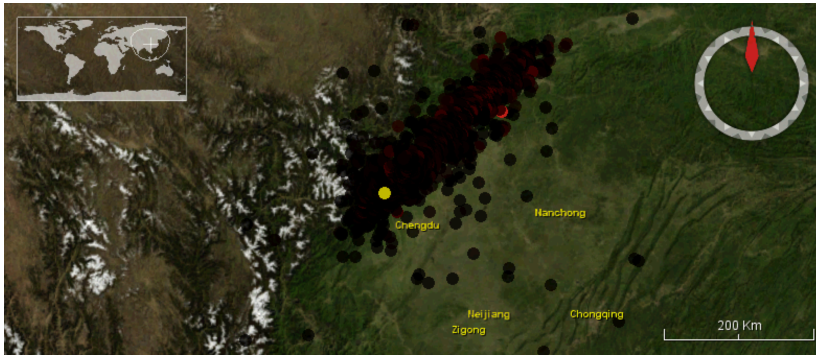




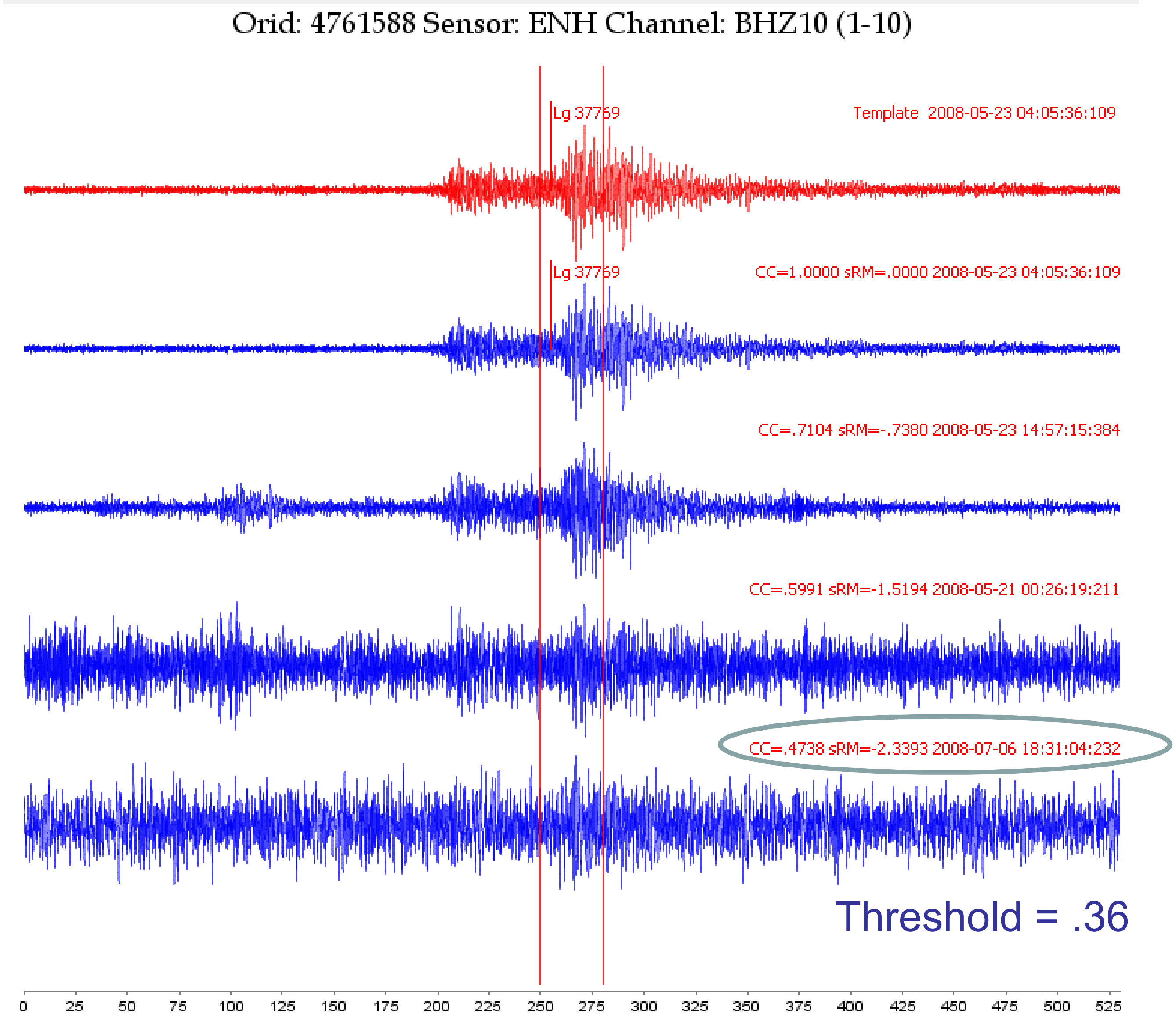
Low threshold (.28)

- Found 17 detections





Detection family at ENH



Slinkard, Megan, Stephen Heck, David Schaff, Nedra Bonal, David Daily, Christopher Young, and Paul Richards "Detection of the Wenchuan Aftershock Sequence using Waveform Correlation with a Composite Regional Network." *Bulletin of the Seismological Society of America* (MS# BSSA-D-15-00333, submitted November, 2015)

- Template approaches – selecting templates, setting thresholds
- **Validating detections**
- Our computational infrastructure
- Datasets and results
  - Central Asia – broad-scale monitoring
  - Wenchuan – replicating local performance with regional stations

Challenges of processing continuous waveform data

- Have 20 (or 40, or 100) opportunities a second to make a false detection!
- 1000 templates * 1 FA/yr = 1000 FA year
- False detections can be on a scale equal to true seismicity
- Need approaches to validate detections

Validation using Multi-Station Confirmation

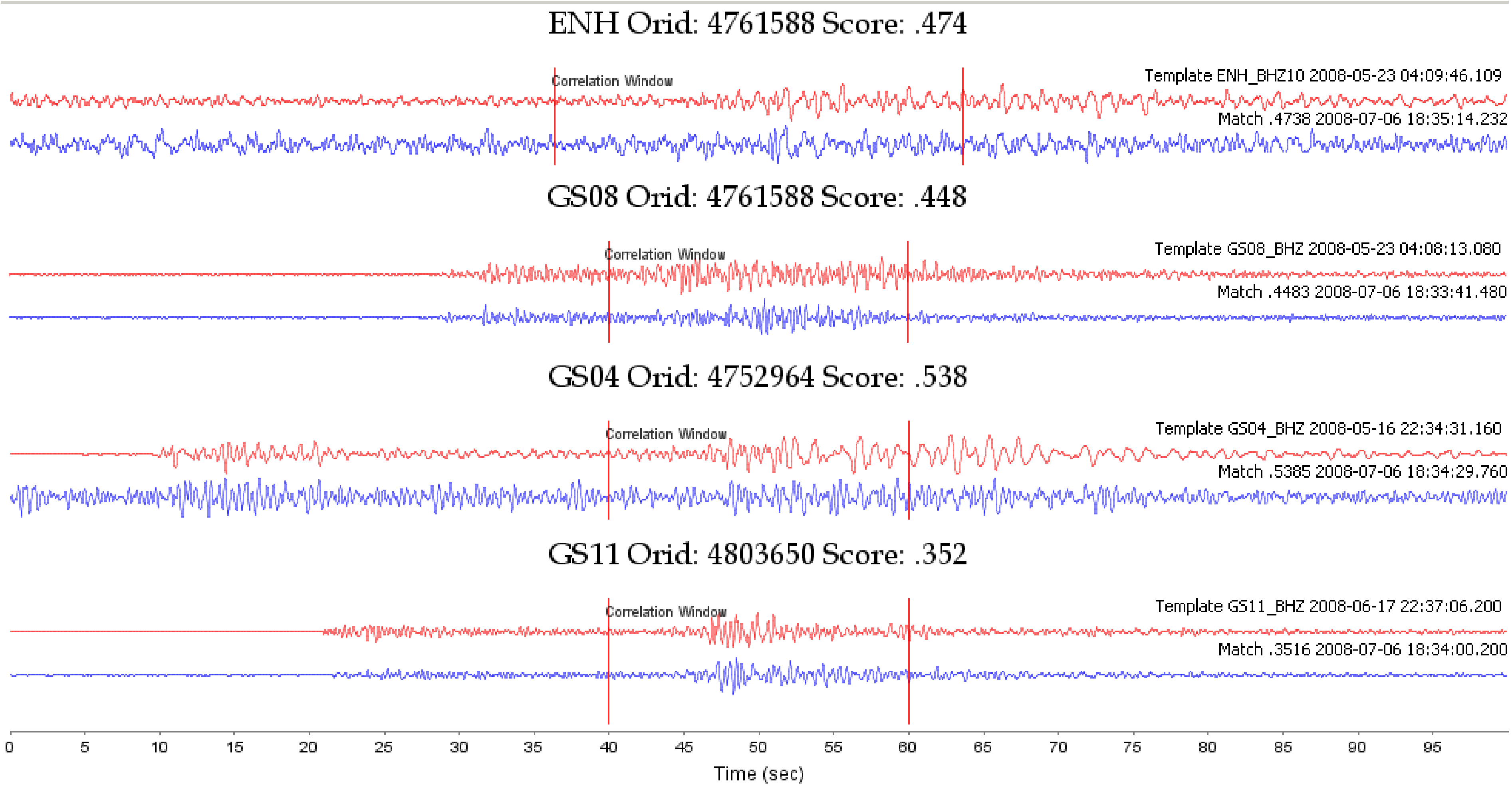
- Probability of more than one station detecting a false event at the same time is very small.
- Use calculated origin times. Require calculated origin times within 4 sec.
- If you have 1000 templates: ~1 false 3 station confirmation every 62,000 years.

Require detecting templates to be geographically co-located

How did this work?

- Central Asia: Between 57% and 71% of the detections at each station were validated by another station
  - Increased to between 66% and 75% validated when include KNDC, LEB, and high correlation as validating means

Detection, as seen at 4 stations



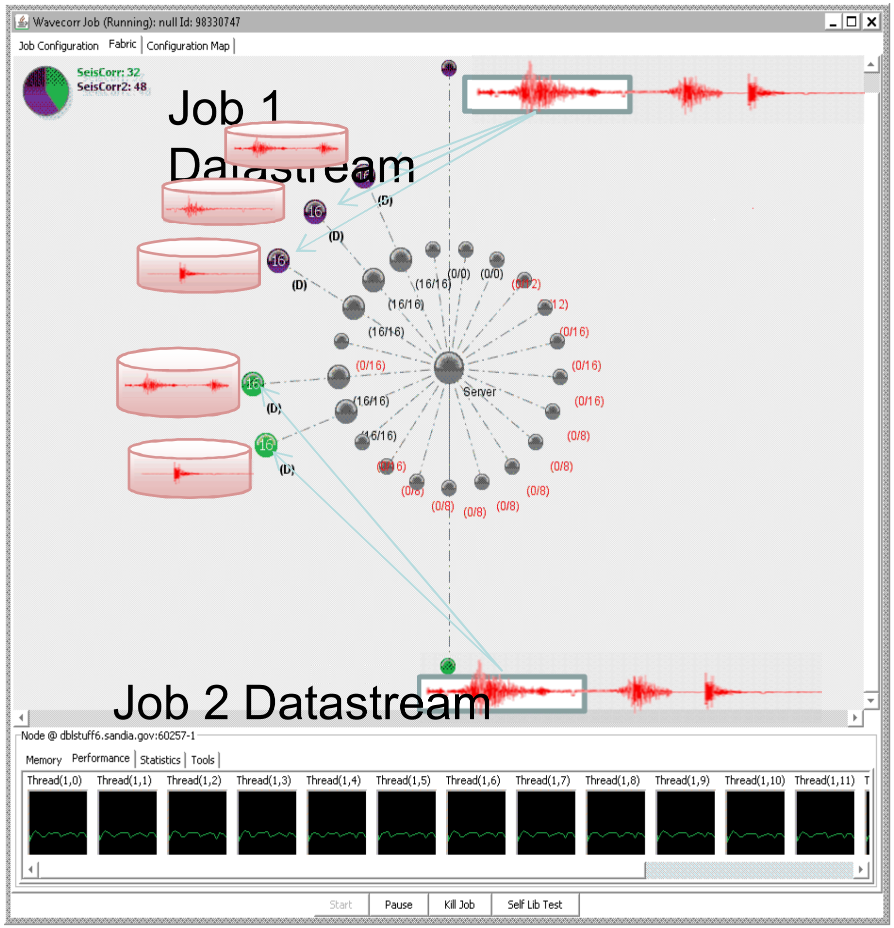
- Template approaches – selecting templates, setting thresholds
- Validating detections
- **Our computational infrastructure**
- Datasets and results
  - Central Asia – broad-scale monitoring
  - Wenchuan – replicating local performance with regional stations

Seiscorr

- Builds template libraries, processes data, validates results
- Distributed parallel system
  - Uses **Fabric**
  - Access to 528 cores
- Computationally efficient algorithms
  - Normalized Cross- Correlation done in frequency domain
- Reads and write to database
- ~400x faster than real time

Seiscorr processing

- Each station processed as its own job
- *Templates* distributed across processing cores
- Raw waveform data processed in ~12 minute chunks; sent to cores for correlation with template group.
- Each core reports back highest correlation score from best template match
- Results integrated
- Find correlation peaks
- Declare detections (template and correlation score)



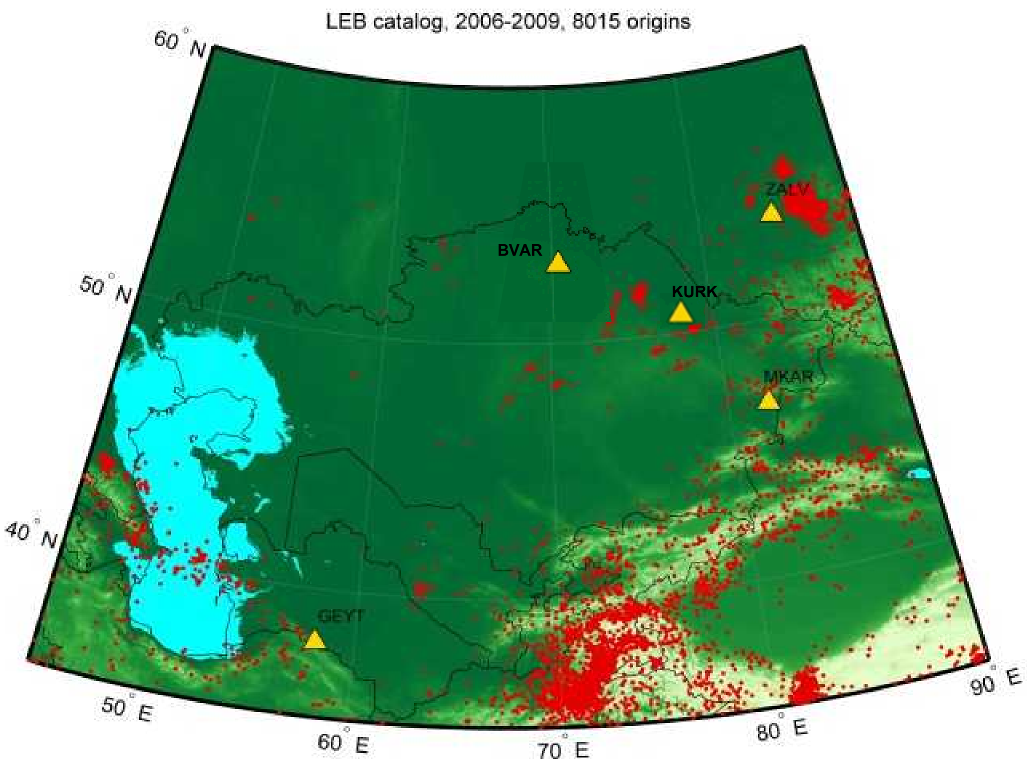
SeisCorr Usage

- Due to overhead of transmitting waveforms, more cores isn't always better. For this work: ~48 cores is optimal.
- Multiple jobs can process at the same time
- 3 stations, 3 years, ~1000 templates a station => 2.5 days

- Template approaches – selecting templates, setting thresholds
- Validating detections
- Our computational infrastructure
- **Datasets and results**
  - Central Asia – broad-scale monitoring
  - Wenchuan – replicating local performance with regional stations

Broad Regional Monitoring

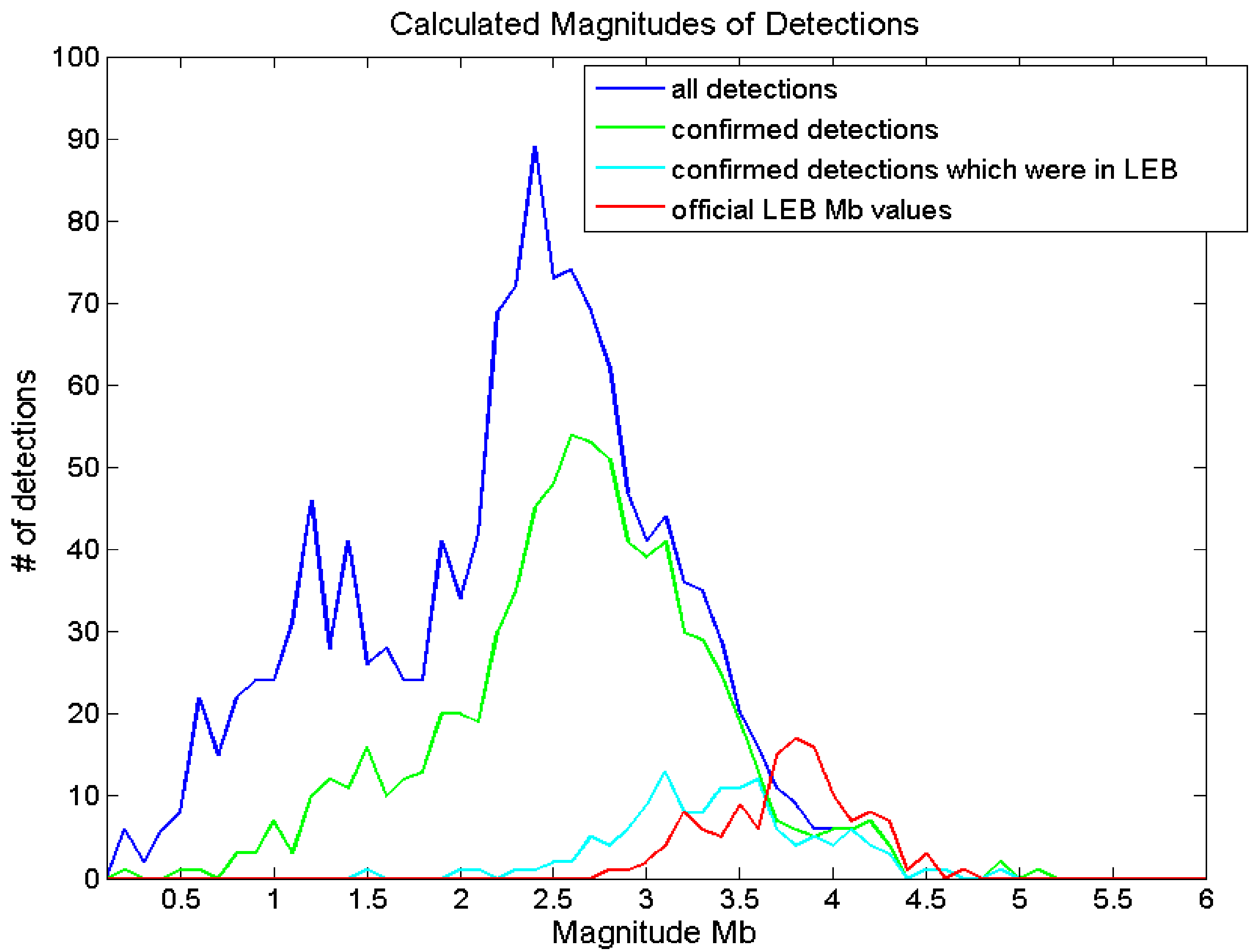
- 3 stations (MKAR, BVAR, KURK)
- Central Asia – chose templates from lat: 35°-60°, lon: 45°-90°
- Lg templates were screened (high STA/LTA) and clustered
- 3 years (2006-2008) of continuous waveform data



Results

- Detected and confirmed 6563 unique events
  - 2+ stations, or in KNDC or LEB bulletin
  - 1481 3 station confirmations
- Detected ~20% of the LEB bulletin
  - 976 detections were in the LEB bulletin
    - Templates were drawn from that same time period…
  - 4886 events were NOT in the bulletin

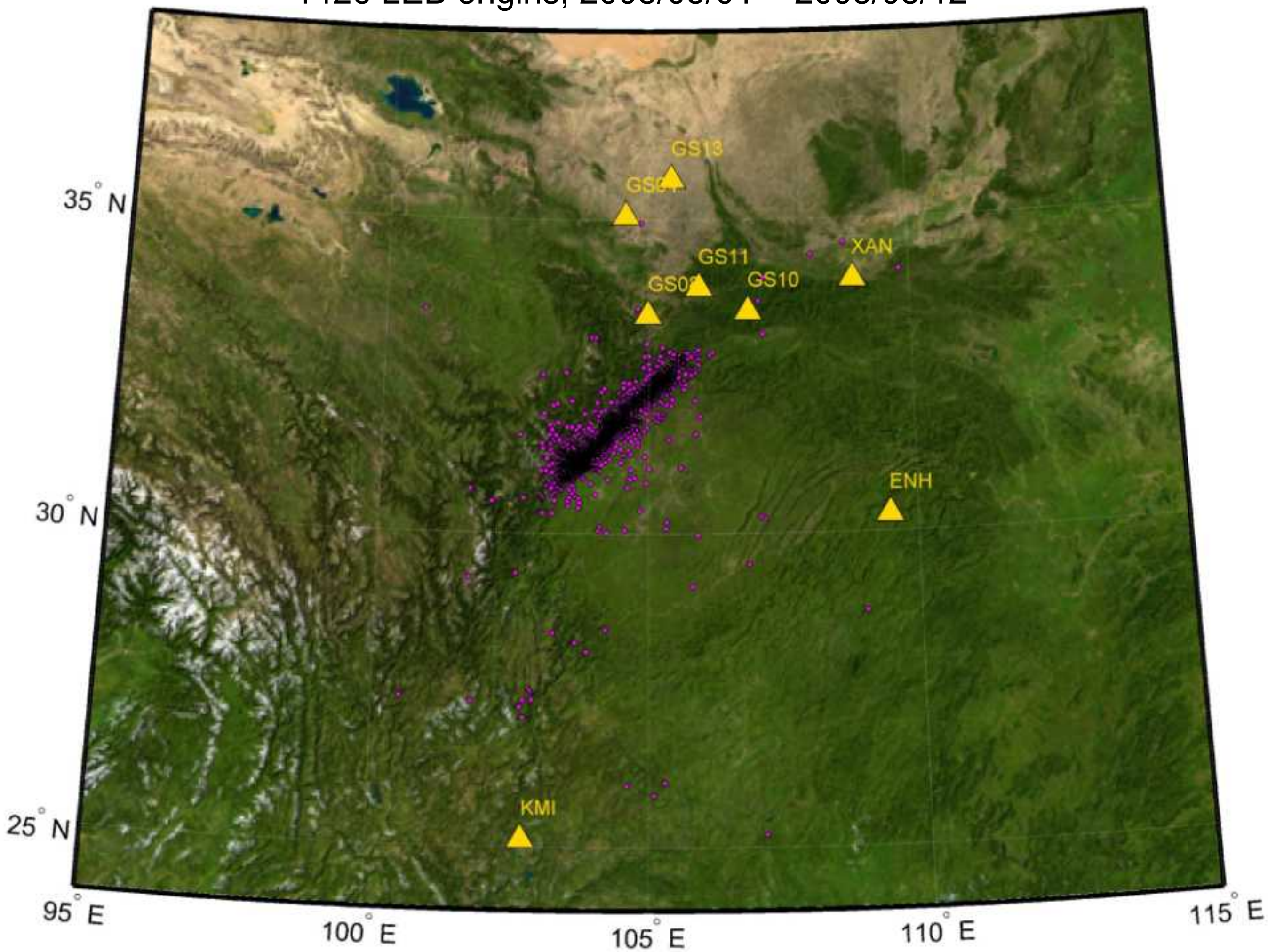
Magnitude of Detections



Slinkard, Megan, David Schaff, Natalya Mikhailova, Stephen Heck, Christopher Young, and Paul G. Richards. "Multistation validation of waveform correlation techniques as applied to broad regional monitoring." *Bulletin of the Seismological Society of America* 104, no. 6 (2014): 2768-2781.

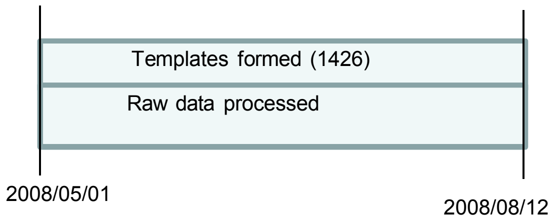
Wenchuan study

1426 LEB origins; 2008/05/01 – 2008/08/12



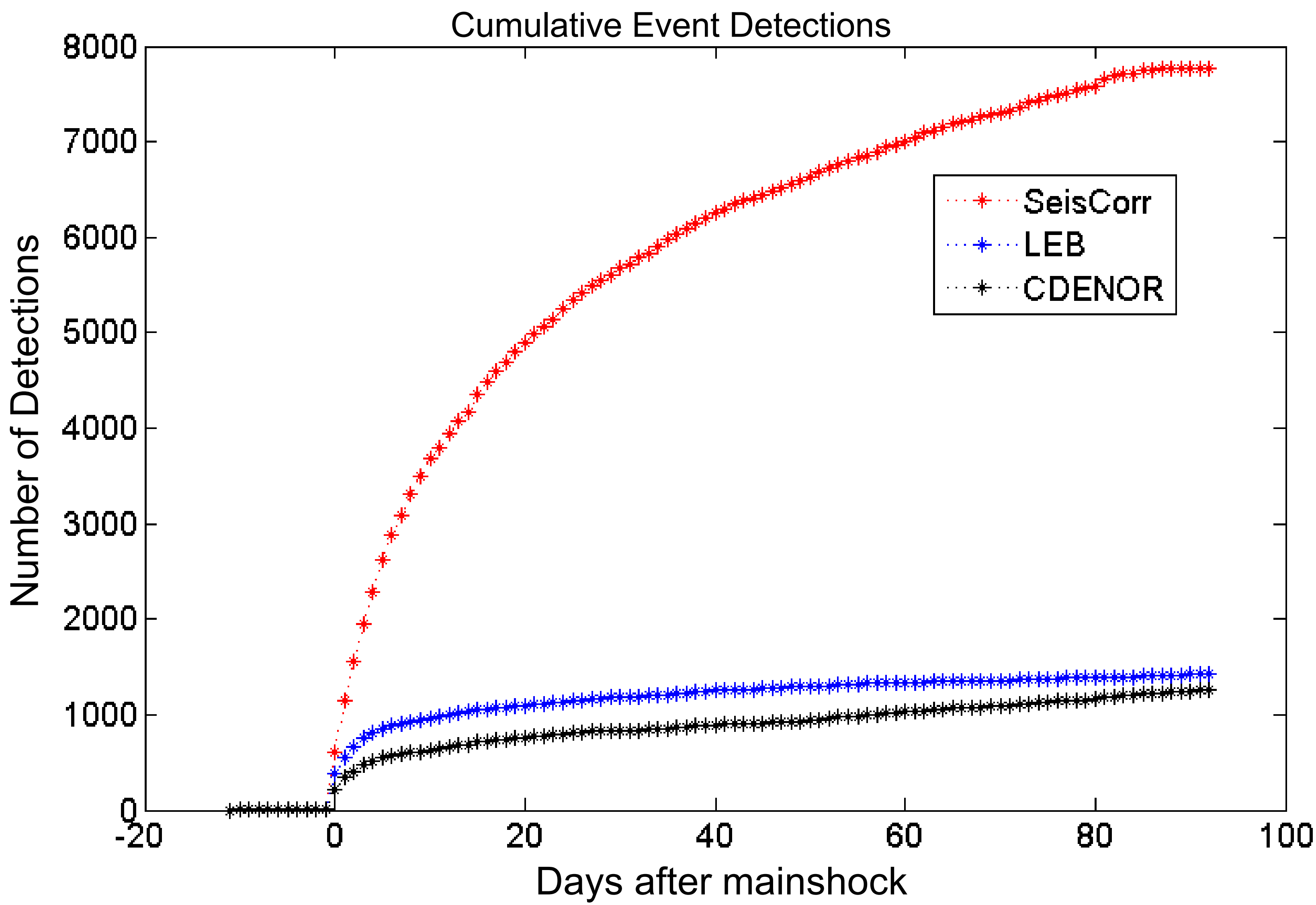
Slinkard, Megan, Stephen Heck, David Schaff, Nedra Bonal, David Daily, Christopher Young, and Paul Richards "Detection of the Wenchuan Aftershock Sequence using Waveform Correlation with a Composite Regional Network." *Bulletin of the Seismological Society of America* (MS# BSSA-D-15-00333, submitted November, 2015)

Experiment Setup

Stations used	XAN, EHN, KMI (IRIS) GS04, GS08, GS10, GS11, GS13 (ASCENT)
Components used	3 (BHZ, BHN, BHE)
Templates: Lat/Lon Box	Lat: 25N – 35N; Lon: 95E – 110E
Template Events: reference bulletin	IDC LEB
Template waveforms	20 sec showing Lg arrivals .5-5Hz bandpass filtering
Dates SeisCorr Ran 	All events in the LEB bulletin between 5/1 and 8/12 were used as template events.

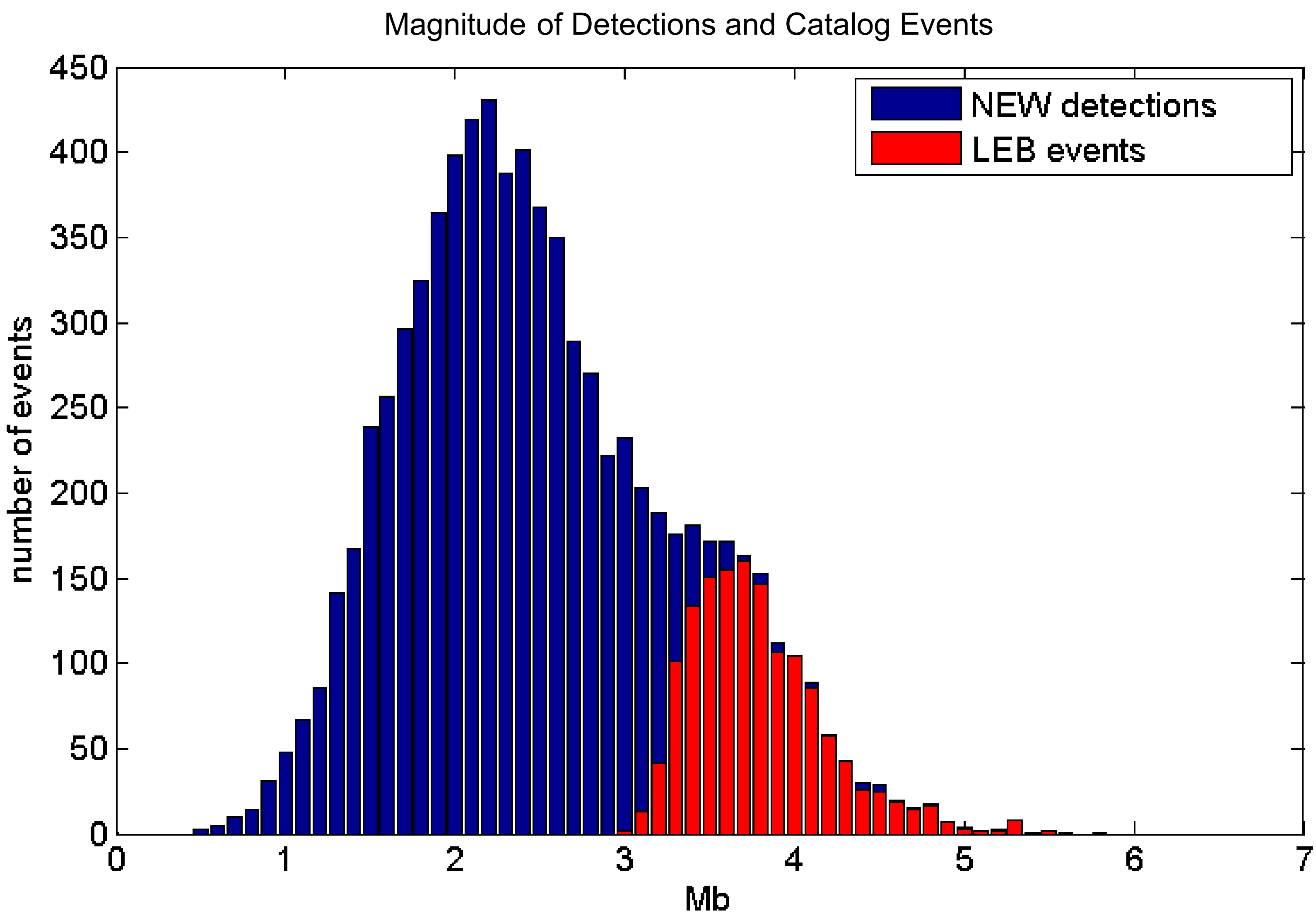
- Because the LEB was not based on any of the stations we processed, we had no arrival pick information
- Assumed that each event had a useful template at each station

Detections of Wenchuan Aftershocks



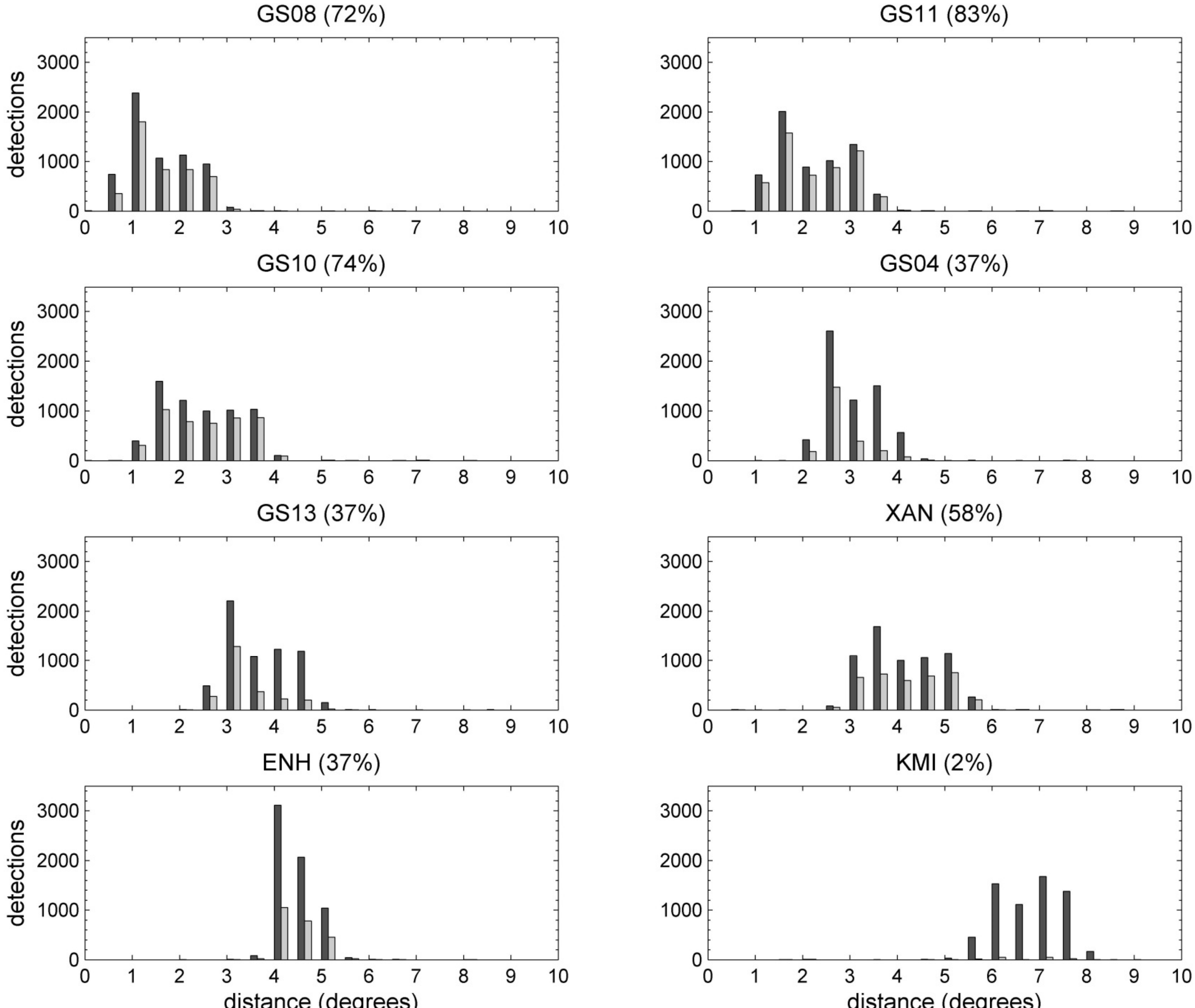
Slinkard, Megan, Stephen Heck, David Schaff, Nedra Bonal, David Daily, Christopher Young, and Paul Richards "Detection of the Wenchuan Aftershock Sequence using Waveform Correlation with a Composite Regional Network." *Bulletin of the Seismological Society of America* (MS# BSSA-D-15-00333, submitted November, 2015)

Magnitude of New Detections



Slinkard, Megan, Stephen Heck, David Schaff, Nedra Bonal, David Daily, Christopher Young, and Paul Richards “Detection of the Wenchuan Aftershock Sequence using Waveform Correlation with a Composite Regional Network.” *Bulletin of the Seismological Society of America* (MS# BSSA-D-15-00333, submitted November, 2015)

Contribution of each station: Network (dark), Validated detections (light)



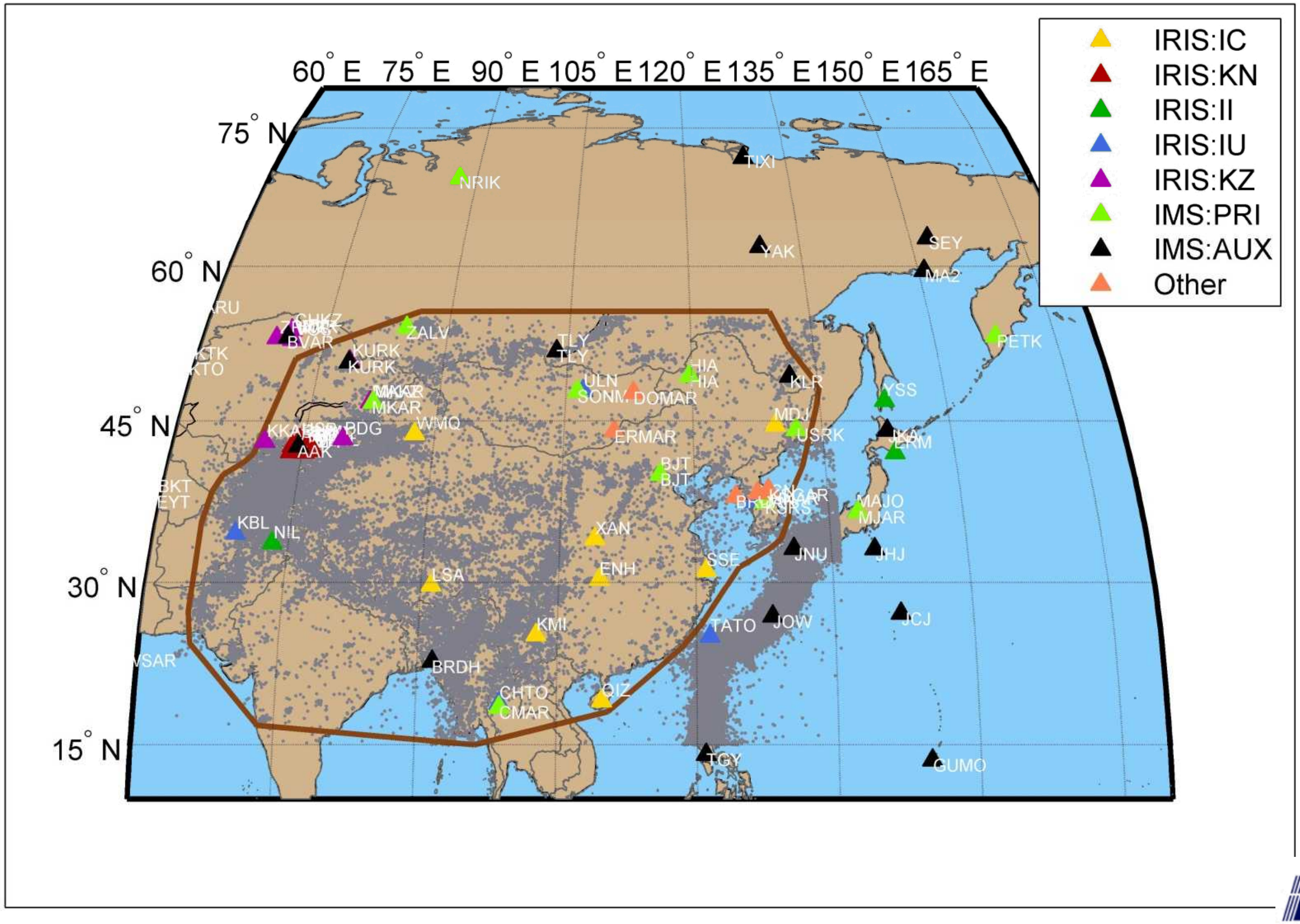
Summary

- Motivation: aid analysts as we search for ever smaller events
- Moving from focused studies to automated processing of continuous waveform data
- Method for setting thresholds which gives consistent FAR across all templates at a station
- SeisCorr allows rapid processing at research speed
- Central Asia work shows feasibility of broad-scale monitoring
  - Detected 20% of LEB
  - Confirmed detections would double size of bulletin
- Wenchuan work shows effectiveness of monitoring using regional stations

Future work

- Monitoring of eastern China using templates going back to time of station install

ISC 1986-2012 155023 events



Backup slides

- text

Other questions to ask

- What percent of the events that make up the current bulletin can you detect using WC?
  - Probably 15-20% during "normal" times, 40-80% during aftershock sequences (Slinkard 2011, 2014), Dodge (2012, 2016)
- What about the red line (analyst added events)?
  - A greater percentage than the blue line! Probably ~50% (Slinkard SnT2015)

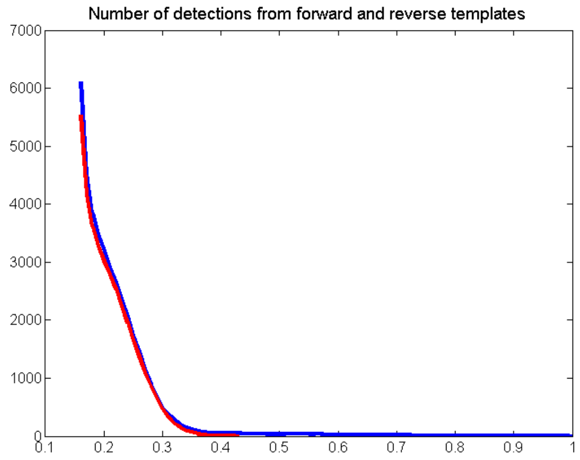
Slinkard, Megan, Stephen Heck, David Schaff, Nedra Bonal, David Daily, Christopher Young, and Paul Richards "Detection of the Wenchuan Aftershock Sequence using Waveform Correlation with a Composite Regional Network." *Bulletin of the Seismological Society of America* (MS# BSSA-D-15-00333, submitted November, 2015)

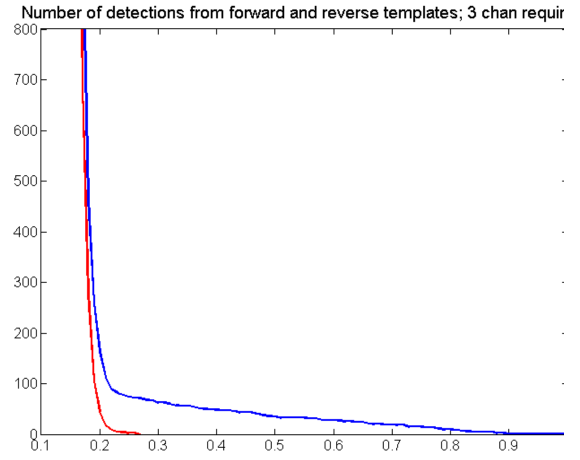
Results Summary

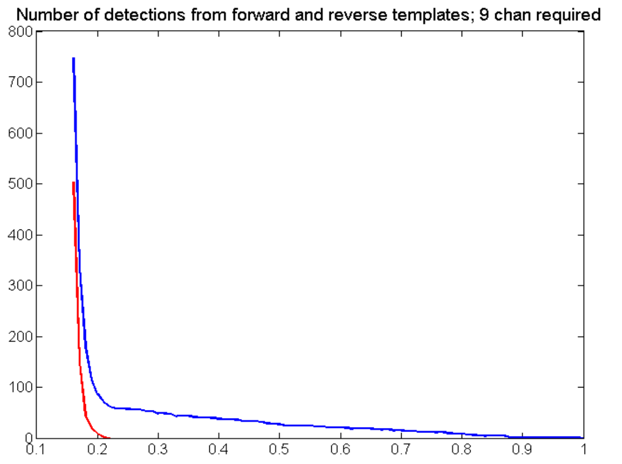
	Number of Detections	Detections confirmed in the LEB	Detections confirmed in the KNDC Catalog	Detections confirmed in either catalog	Detections confirmed by another station (1409 seen by all 3 stations)	Detections confirmed by high correlation (>0.7)	Confirmed in some manner
MKAR (array)	7426 (from 506/812 templates)	526 (10.8% of catalog)	553	927	4740 (64%) (4226 not in LEB)	1309 (183 not in LEB or another station)	5136 (69%) (4610 not in LEB)
BVAR (array)	3096 (from 193 /537 templates)	154 (3% of catalog)	67	180	2199 (71%) (2054 not in LEB)	383 (17 not in LEB or another station)	2307 (75%) (2153 not in LEB)
KURK (3C)	8526 (from 837/1515 templates)	693 (14% of catalog)	705	1179	4895 (57%) (4327 not in LEB)	2031 (376 not in LEB or another station)	5648 (66%) (4955 not in LEB)

Number of detections from Forward and Reverse templates (all of 2006)

We limit detections in times of data dropouts by requiring at least 3 channels to have contributed to the correlation. This impacts the calculated Correlation Threshold significantly.







1 good channel required

3 good channels required

9 good channels required

Questions that need answering before the problem is solved:

How many events can we detect using correlation?

How do we best use templates to detect nuisance repeated events?

How do we best use templates to find interesting repeated events?

How specific are templates? Should we use subspace or EMFD to make them more general?

What are the differences between running this on small scale vs large scale datasets?

How do you set detection thresholds? How do you validate detections?

Can this run in real time? Fast enough for research?

How should we apply waveform correlation across a network?

How do you get this working smoothly in an operational system?