This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

SAND2023-04779C

Testing a Paired Neural Network to Characterize Aftershock Sequences

Erica Emry

Brendan Donohoe, Andrea Conley, Rigobert Tibi, and Christopher Young

June 15, 2023

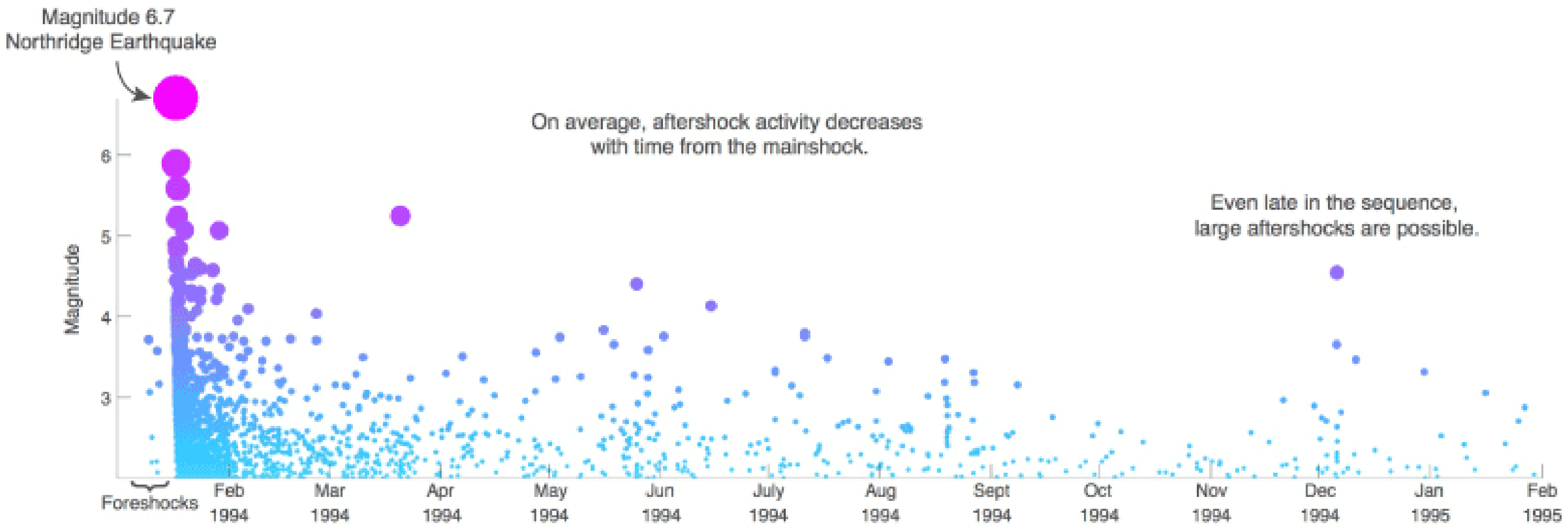
This Ground-based Nuclear Detonation Detection (GNDD) research was funded by the National Nuclear Security Administration, Defense Nuclear Nonproliferation Research and Development (NNSA DNN R&D).

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Aftershock Sequences and Cross-Correlation

- Large-magnitude earthquakes & aftershock sequences unexpectedly occur & greatly increase analyst workload
- Cross-correlation techniques can identify similar earthquakes (like aftershocks)
  - Creation of quality template libraries for in progress sequence can be difficult
  - Some regions have no historical seismicity to use as templates
  - Cross-correlation can be affected by spikes and overlapping earthquakes
  - Cross-correlation tends to be computationally intensive



From: USGS Aftershock Forecast Overview
https://earthquake.usgs.gov/data/oaf/overview.php

Can ML Improve Aftershock Labeling?

- Like cross-correlation, if similar events (such as aftershocks) can be rapidly labelled a match or non-match, then could help alleviate analyst burden
- Allows analyst to maintain their attention on other events globally

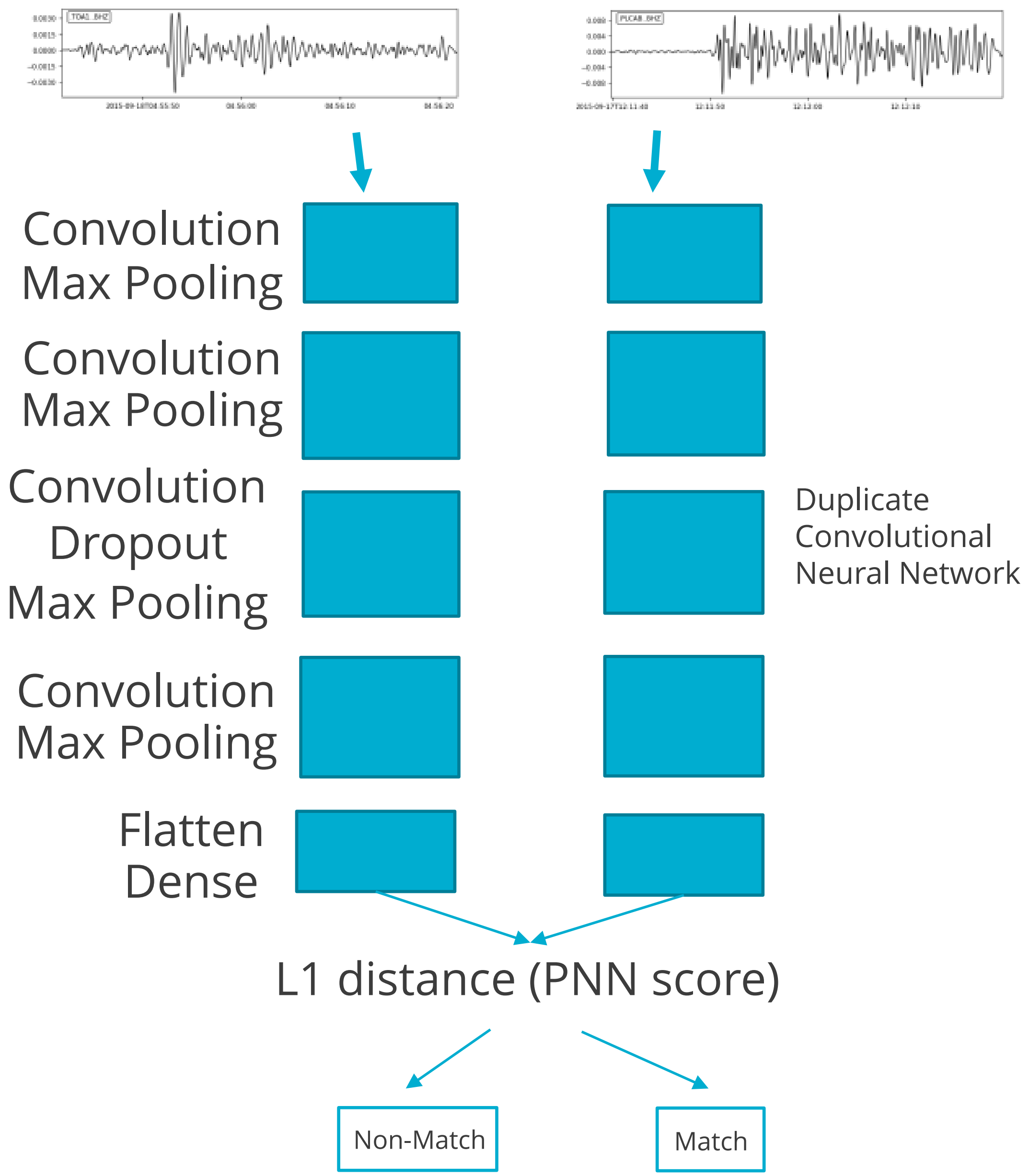
- Any ML model *must* generalize to data from other regions that it wasn't trained on.
  - Location of the next large magnitude earthquake is unpredictable and can occur around the globe.
  - We may have few records of seismicity on faults that could produce a large-magnitude earthquake (currently locked faults, like Cascadia subduction zone)
- A ML model should not mislabel events we care about as an aftershock. The model should have low likelihood of false positives (false classification as aftershock).

Paired Neural Network

Conley et al. (2021) trained a PNN model to identify waveform similarity

2 branches with same architecture:

- Each branch is a Convolutional Neural Network (CNN)
- Branches are exact duplicates
- 4 blocks with 2 or 3 transformations in each
  - Convolution
    - Some # of convolutional filters
    - Size of the output is not altered
  - Max Pooling
    - Takes the top number in each group
    - Size of the output decreases
  - Spatial Dropout
    - To prevent overfitting
- Monte Carlo dropout used to quantify uncertainty



Initial Training Data
Real Event Data, with added Noise

Trained on global seismicity

Not aftershock earthquakes

- 15,764 earthquakes
- 827 stations

Global distribution of stations:

- Much denser in U.S., Europe, Australia, and East Asia

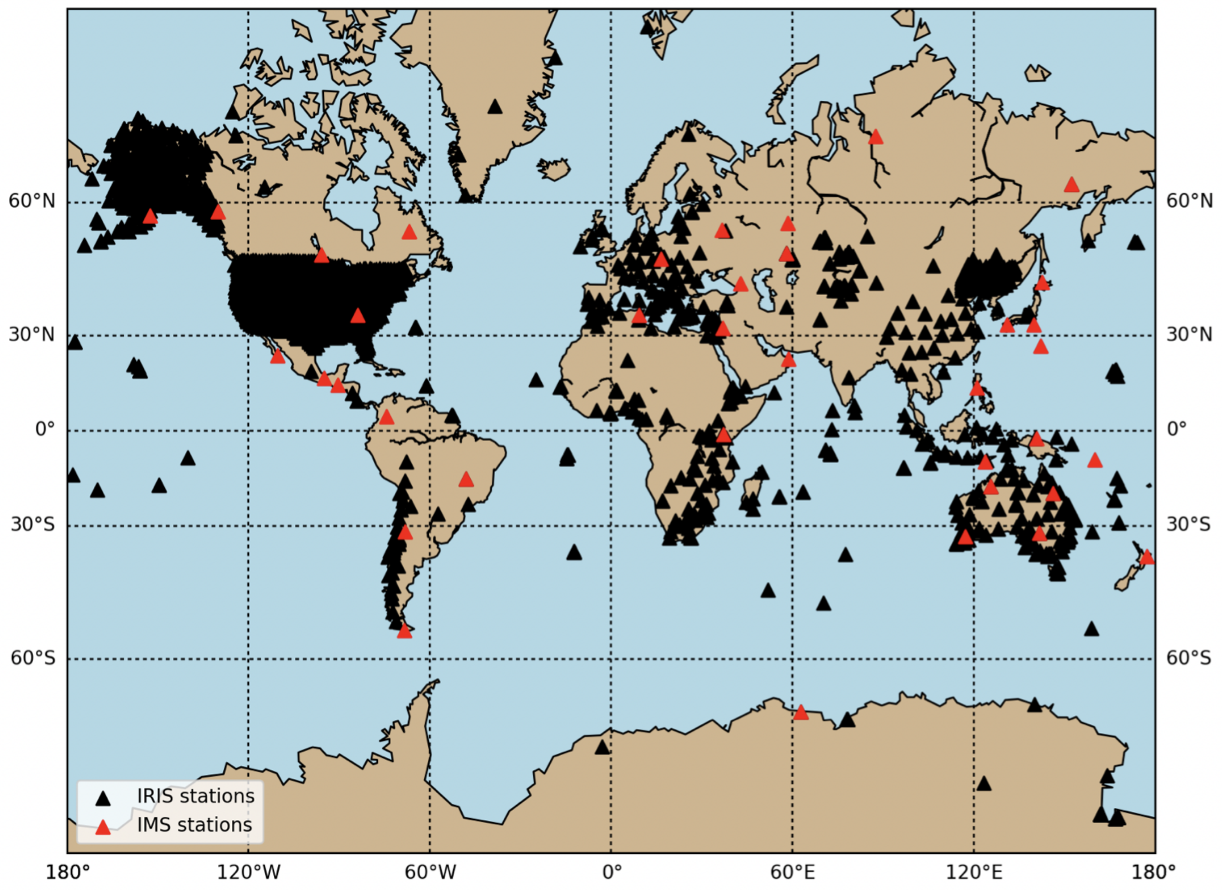
Noise datasets

- STEAD noise dataset (Mousavi et al., 2019)
- University of Utah noise dataset (Tibi et al., 2021)

Some training datasets included ‘overlapping’ waveforms

Training datasets were filtered at different frequencies

- Raw, bandpassed (1.5-5 Hz), highpassed (>0.3 Hz)

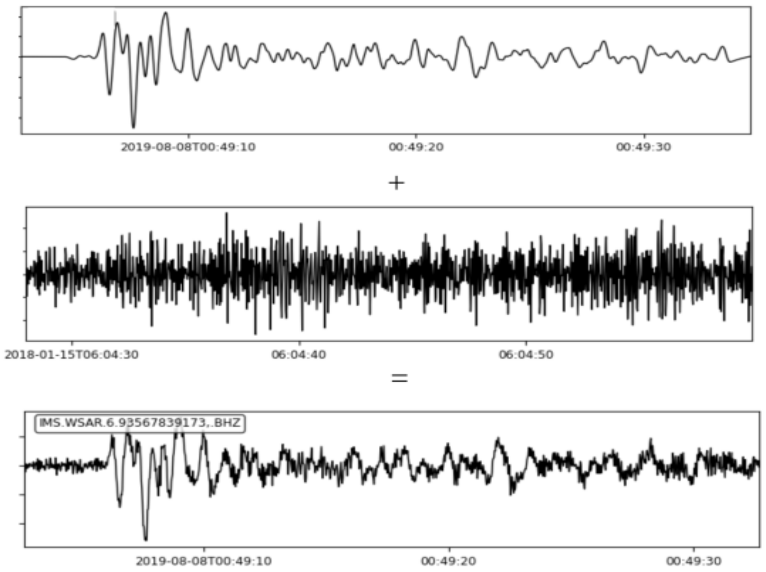


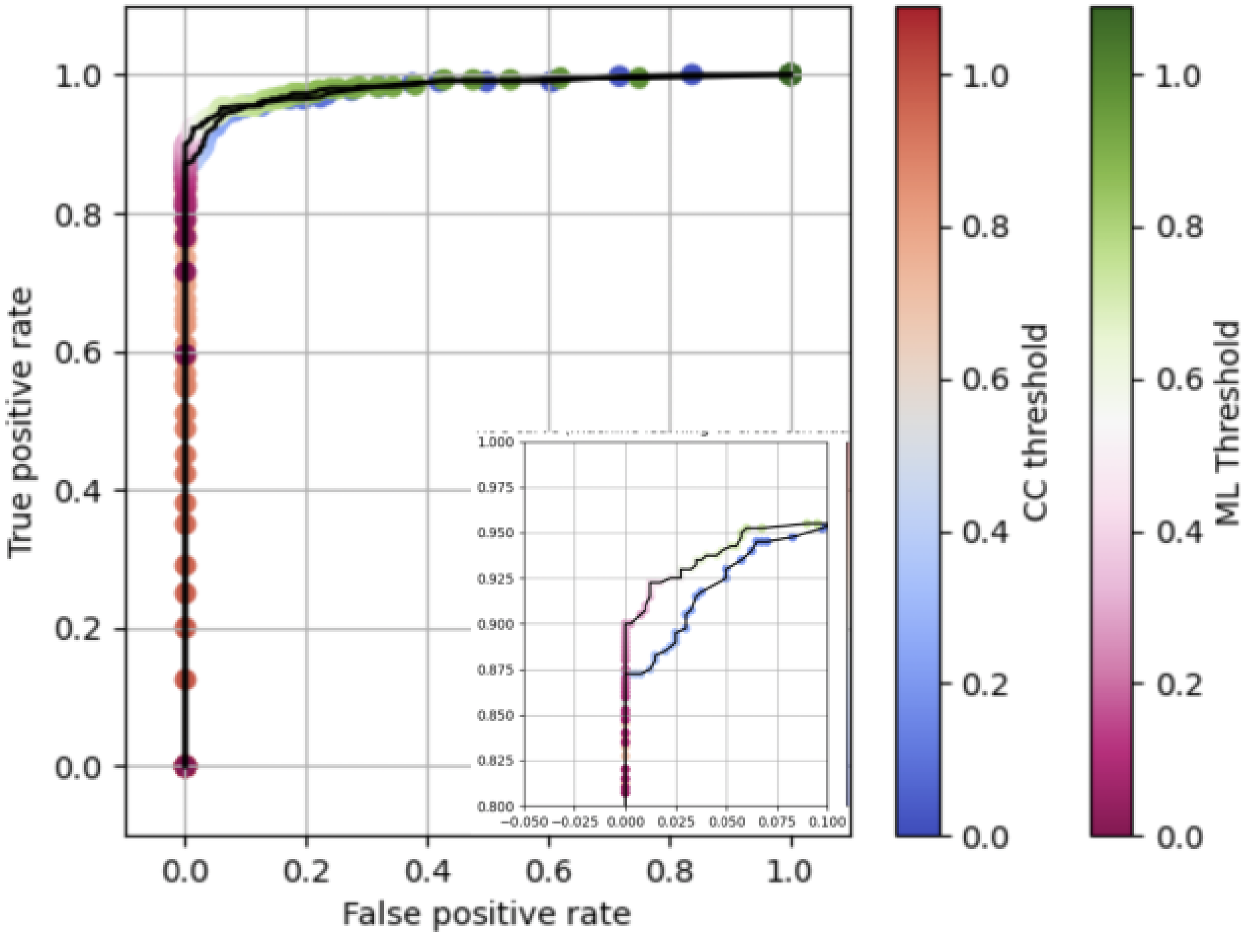
Conley et al., (2021) Training Data Station Distribution

Original PNN model test results

Tested against subset of constructed data (15%):

- Test data randomly pulled out of full dataset prior to training
- Outperformed cross-correlation in the top left corner of the ROC curve (magnified in bottom right).



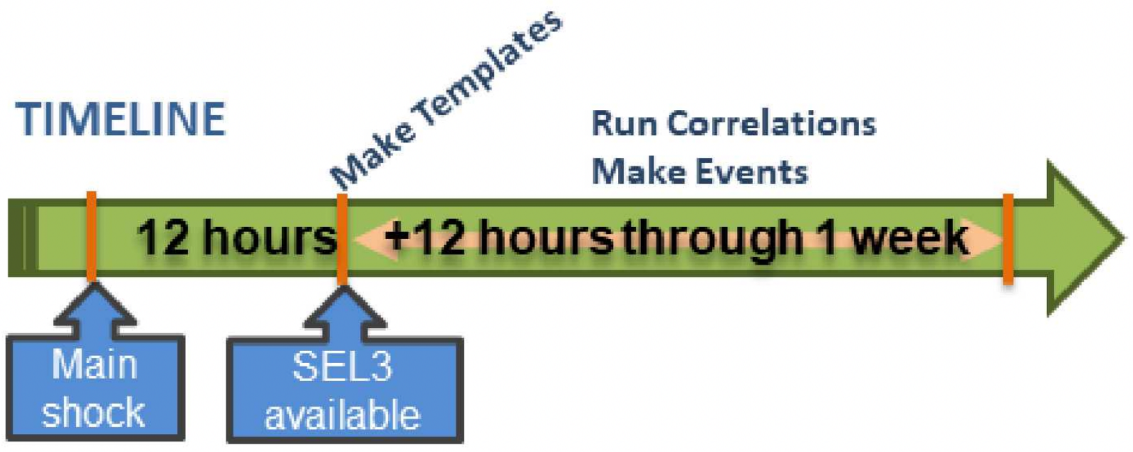


Conley et al. (2021) – Above: ROC curve comparison w/CC scores, Left: Constructed waveform

Test Aftershock Dataset

Aftershock Sequences:

- 2015 Illapel, Chile; 2015 Gorkha, Nepal
- Aftershocks originally from cross-correlation project, using templates from SEL3 automated detection (Sundermier et al., 2019)



Sundermier et al. (2019) - Timeline

Analyst Validation:

- ”True Positive”: arrival matches LEB
- “Valid Added”: valid arrival, not in LEB
- “False Alarm”: arrival not from a valid event or from non-aftershock earthquake elsewhere

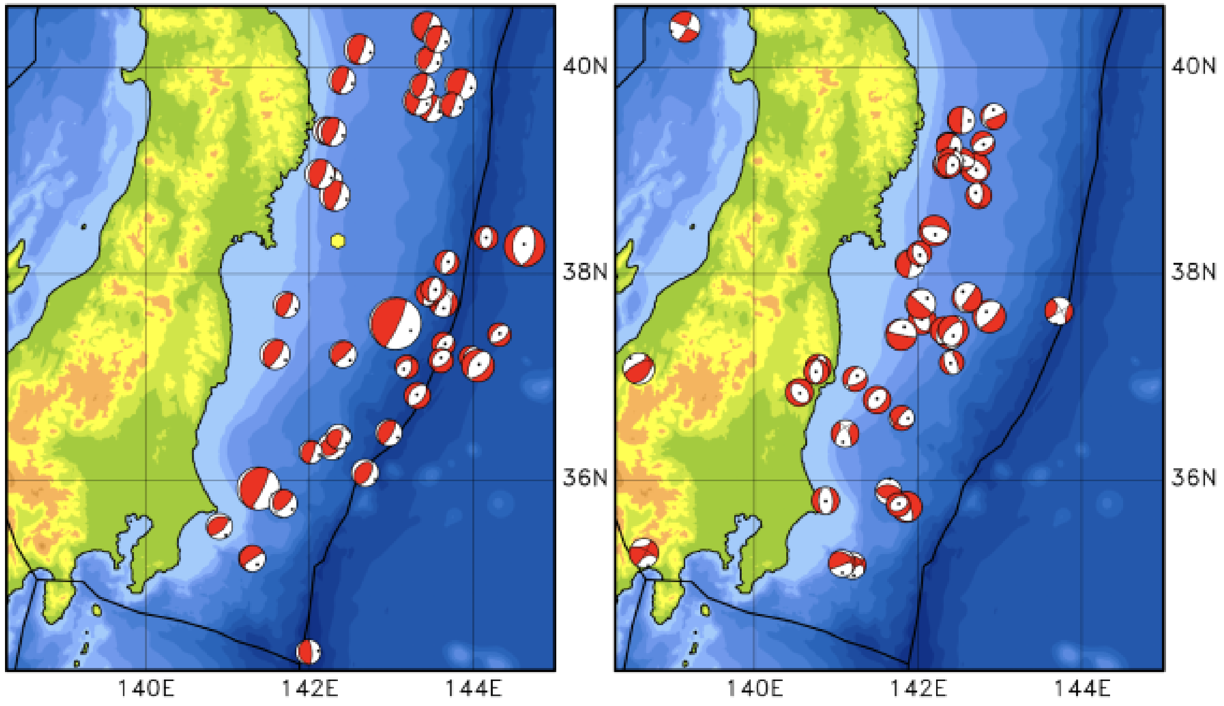
What We Have To Work With:

- 6 different PNN models
  - Trained with or without overlapping signal
  - Trained with data filtered at different frequencies
    - Raw, Bandpass Filtered (1.5-5 Hz), and Highpass Filtered (>0.3 Hz)
- 2 Validated Aftershock Sequences
  - 2015 Illapel, Chile
    - Recorded on vertical component of 12 IMS stations
  - 2015 Gorkha, Nepal
    - Recorded on vertical component of 13 IMS stations
  - Info about ‘True Positives’ and ‘False Alarms’

Match vs. Non-Match Criteria

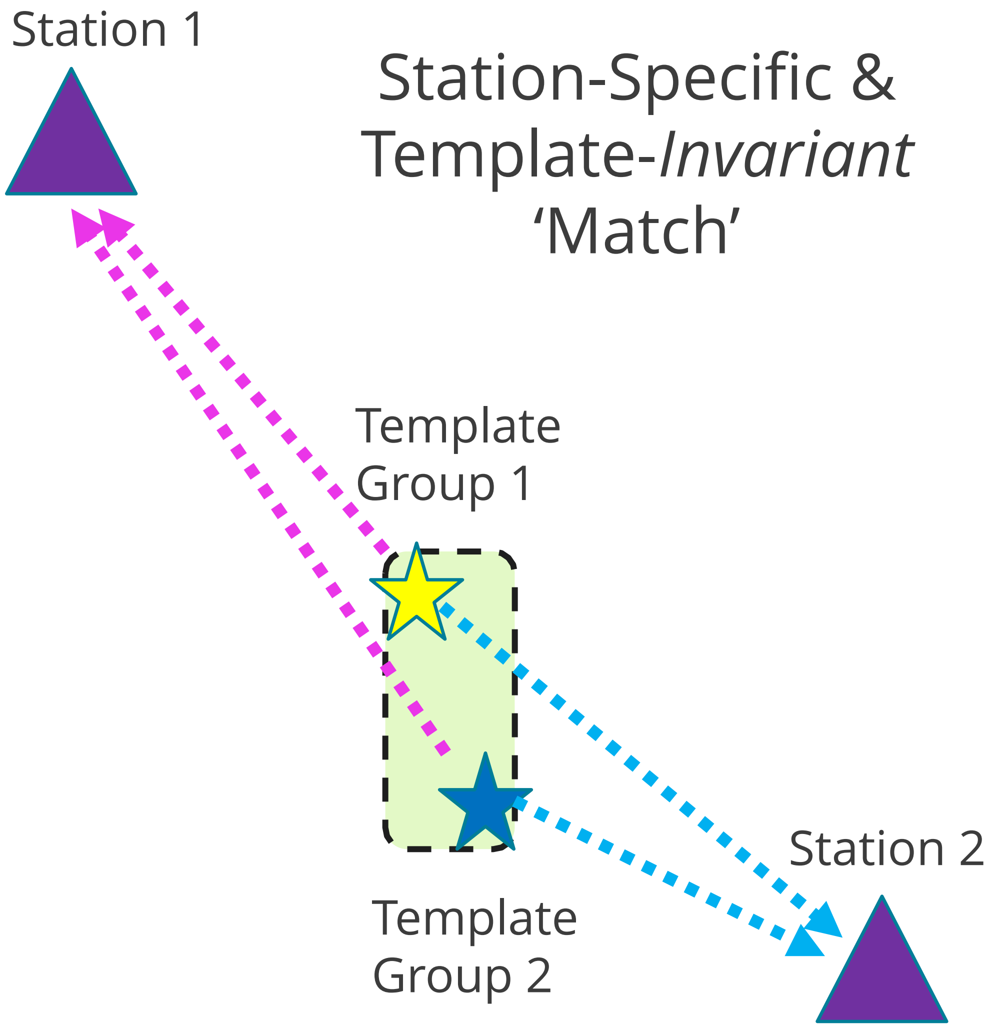
Criteria 1: Event is a match if it is an aftershock.

- What about aftershocks of different orientation/slip?
- What about aftershocks from different ends of the ruptured region?



Nettles et al. (2011), “Conforming” and “Non-Conforming” Aftershocks

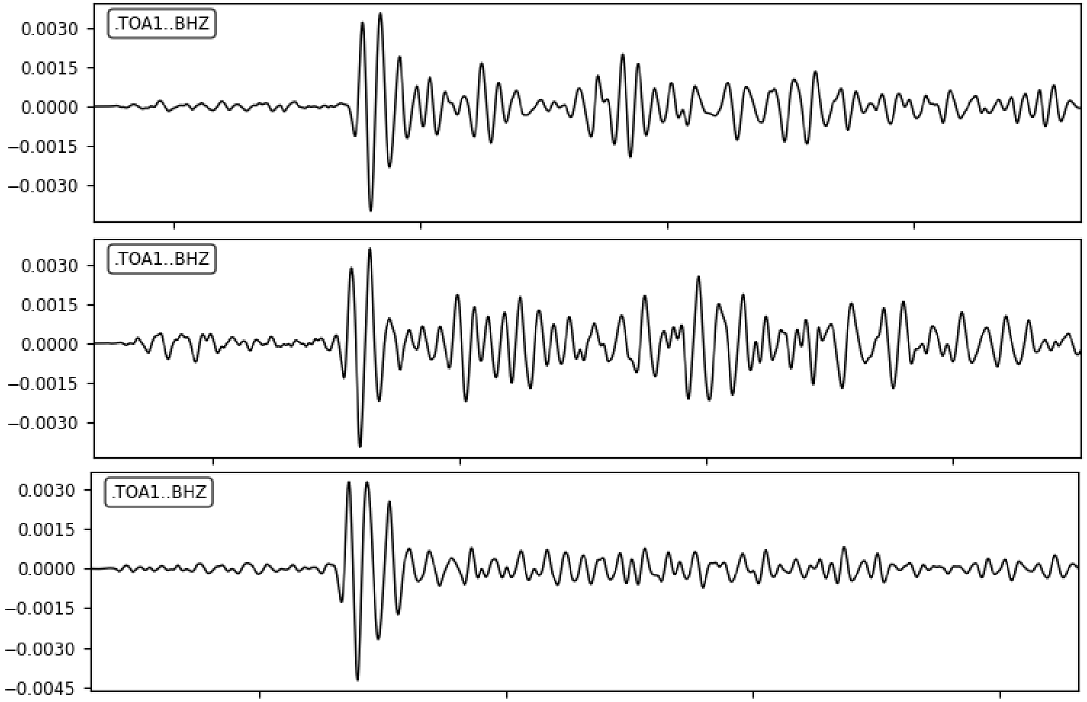
Station-Specific & Template-*Invariant* ‘Match’

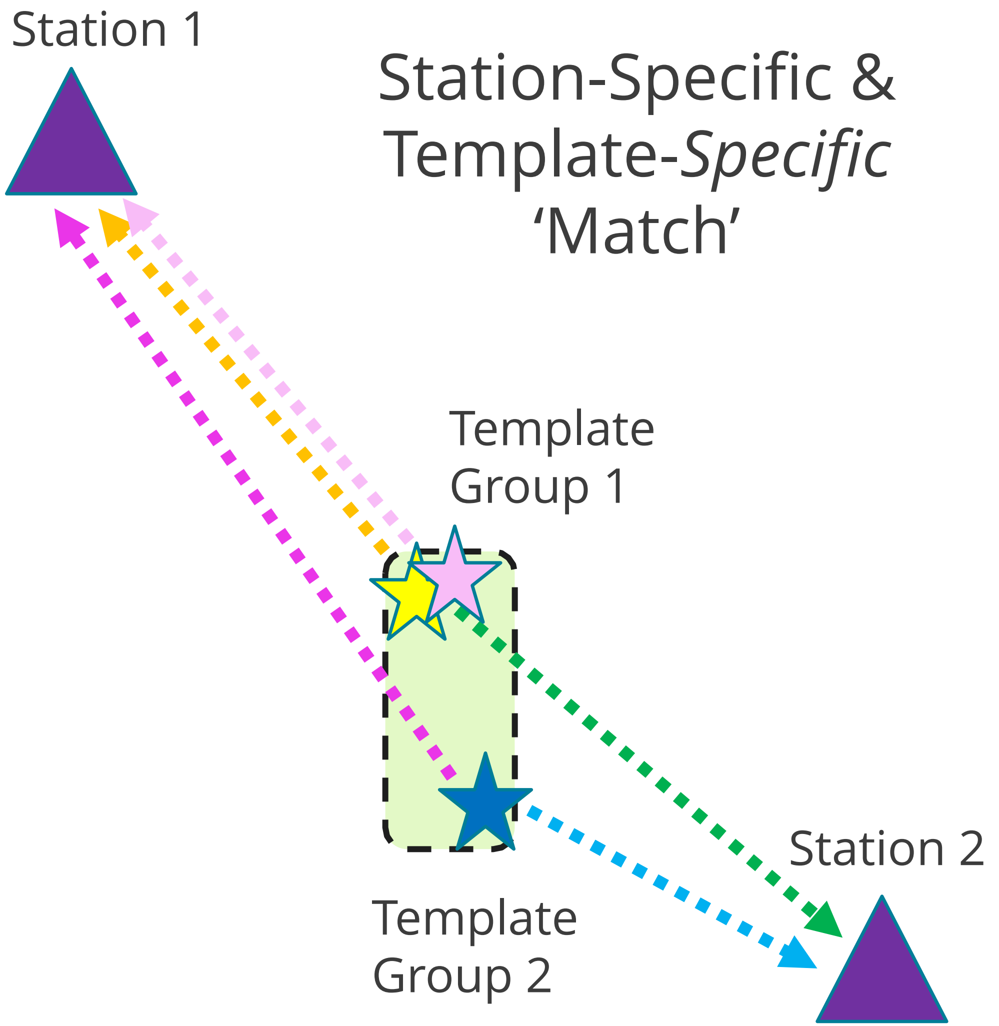


Match vs. Non-Match Criteria

Criteria 2: Event is a match if it was originally detected by the same template event.

- What about events detected by similar template events? Are those a match or non-match?
- This could lead to higher 'false positives'





Results: Statistical Scoring

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

*Evenly sized match & nonmatch populations (no "class imbalance")
This is not true for this test dataset.
More non-matches & model performed best with non-matches
– Accuracy gives artificially high values (making it look better than it is)

$$Precision = \frac{TP}{TP+FP}$$

*Model Precision only sensitive to match predictions
- If our 2nd match criteria ("template-specific") leads to many FP in comparison to the 1st match criteria ("template-invariant"), we'd expect a decreased precision.

$$TPR = \frac{TP}{TP+FN}$$

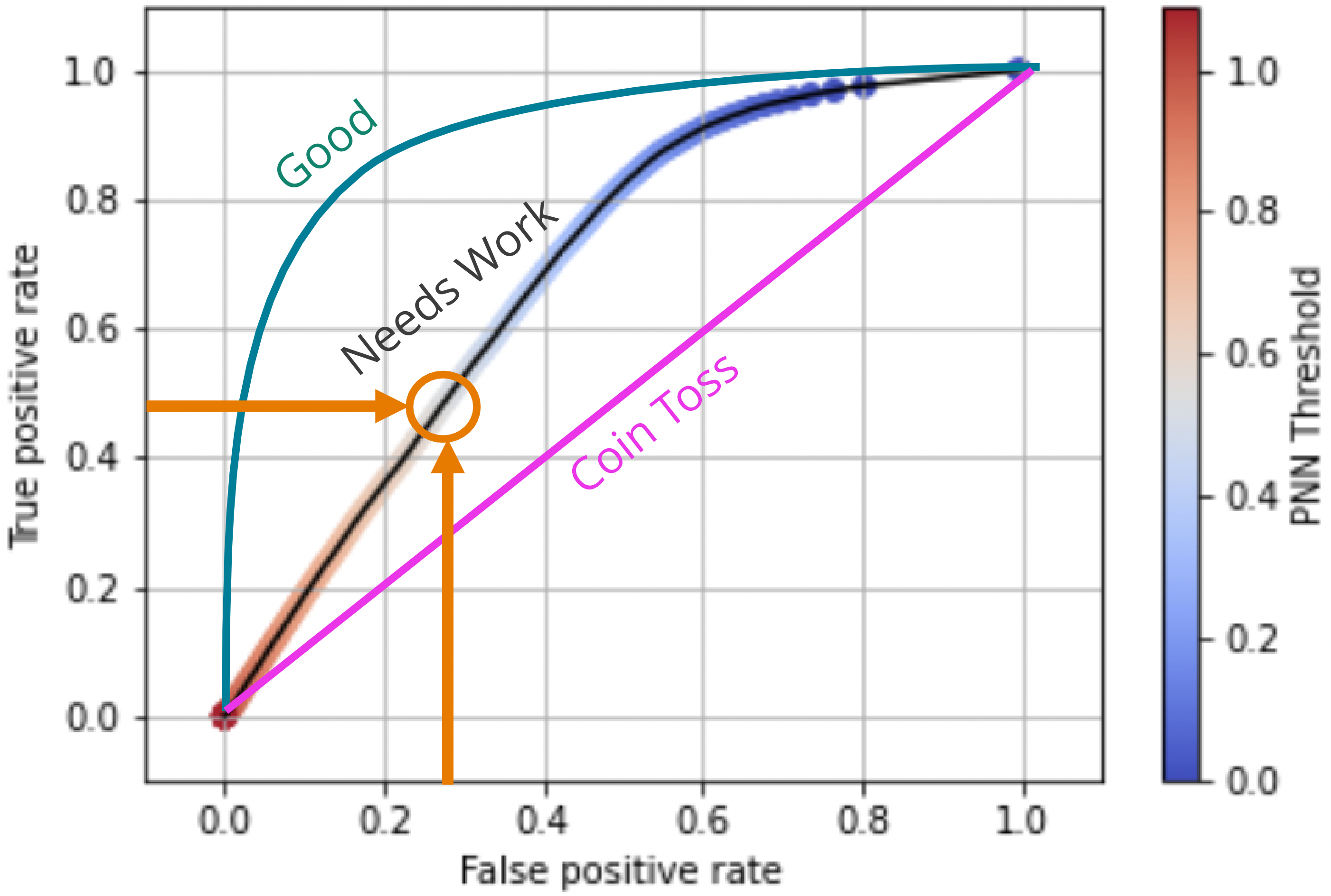
$$FPR = \frac{FP}{TN+FP}$$

*Information for TPR (recall) & FPR with different PNN score threshold assumptions – directly informs the ROC curves

$$F1 = 2 * \frac{Precision * TPR}{Precision + TPR}$$

*Harmonic mean of precision & recall (TPR)
*Specifically designed to handle class imbalance

ROC Curves & AUC



*As PNN score threshold changes (between match & nonmatch), the TPR (Recall) and FPR changes.

*A concave downward curve is desired, and it produces a high Area-Under-the-Curve (AUC)

Coin Toss: AUC = 0.5
Needs Work: AUC = 0.65
Good: AUC ~ 0.9

At a PNN Threshold = 0.5,
the TPR ~0.5 & the FPR ~0.3

Test 5: Bandpass-filtered, Trained without overlapping waveforms, Template-Invariant criteria

Results - Models trained on Overlapping Data

All numbers (except AUC) are for PNN threshold score = 0.5

Match Criteria #1

Station-Specific & Template-Invariant

(All aftershocks are Matches) ☆

Match Criteria #2

Station-Specific & Template-Specific

(All aftershocks associated to similar Templates are Matches) ☆

Test # (Filter)	TPR-to-FPR	AUC	TP	FP	TN	FN	TPR	FPR	Precision	F1
Chile, Station-Specific, Template-Invariant, With Overlapping Signals										
1 (R)	1.20	0.486	606	6738	417899	31247	0.019	0.016	0.083	0.031
2 (BP)	2.05	0.545	2149	13993	410644	29704	0.068	0.033	0.133	0.090
3 (HP)	1.20	0.504	979	10905	413732	30874	0.031	0.026	0.082	0.045
Chile, Station-Specific, Template-Specific, With Overlapping Signals										
7 (R)	1.73	0.515	75	7567	461702	2620	0.028	0.016	0.010	0.015
8 (BP)	3.21	0.587	292	15850	438036	2313	0.112	0.035	0.018	0.031
9 (HP)	1.64	0.528	116	11768	442118	2620	0.042	0.026	0.010	0.016

For PNN Models Trained with Overlaps:

- **Bandpass is better (in comparison to raw or highpass)**
- **AUC is a little higher overall for the template-specific criteria**
- **Clear decrease in Precision in Template-Specific Criteria**

Results - Models trained on NO Overlapping Data

All numbers (except AUC) are for PNN threshold score = 0.5

Match Criteria #1

Station-Specific & Template-Invariant

(All aftershocks are Matches)

Match Criteria #2

Station-Specific & Template-Specific

(All aftershocks associated to similar Templates are Matches)

Test # (Filter)	TPR-to-FPR	AUC	TP	FP	TN	FN	TPR	FPR	Precision	F1
Chile, Station-Specific, Template-Invariant, Without Overlapping Signals										
4 (R)	2.98	0.677	4192	18745	405892	27661	0.132	0.044	0.183	0.153
5 (BP)	1.75	0.688	16869	128390	296247	14984	0.530	0.302	0.116	0.190
6 (HP)	2.22	0.692	4867	29215	395422	26986	0.153	0.069	0.143	0.148
Chile, Station-Specific, Template-Specific, Without Overlapping Signals										
10 (R)	2.69	0.654	349	22588	431297	2256	0.134	0.050	0.015	0.027
11 (BP)	1.66	0.687	1357	143304	310598	1230	0.525	0.316	0.009	0.017
12 (HP)	2.32	0.689	447	33635	420250	2158	0.172	0.074	0.013	0.024

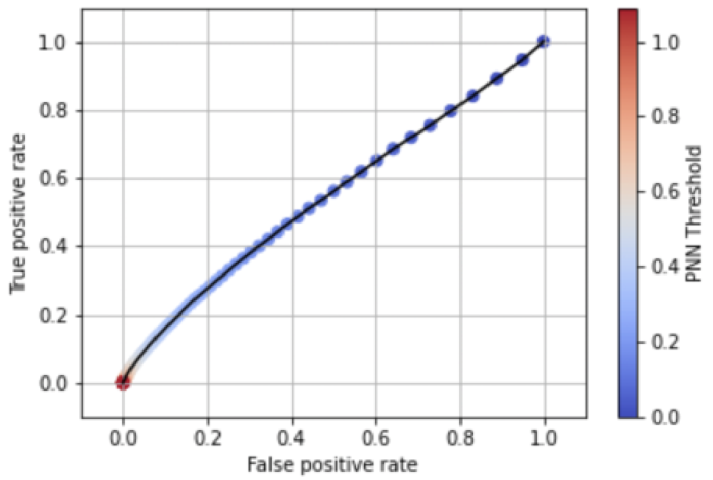
For PNN Models Trained without Overlaps:

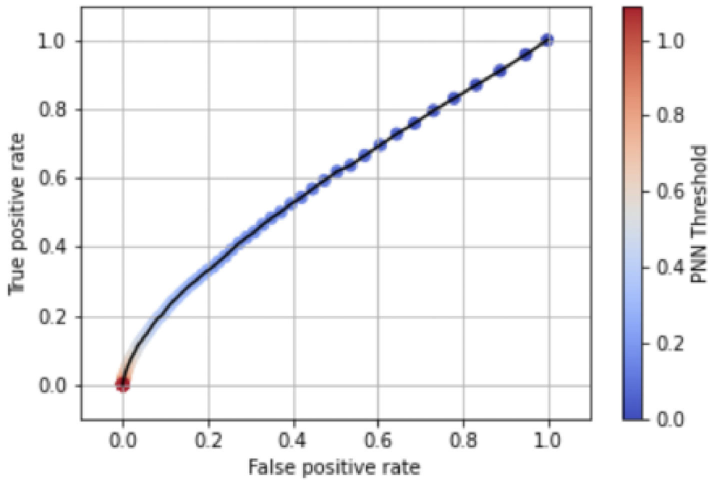
- **AUC is highest overall for these tests**
- **Precision and F1 are highest, assuming Template-Invariant Criteria**
- **Clear decrease in Precision and F1 in Template-Specific Criteria**

ROC Curve Comparisons

- Trained with Overlapping Data – Bandpass Filtered

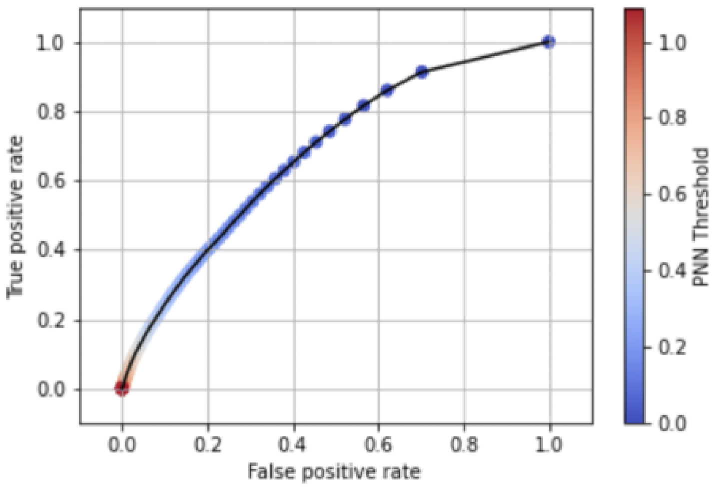
Test 2:
Template-Invariant



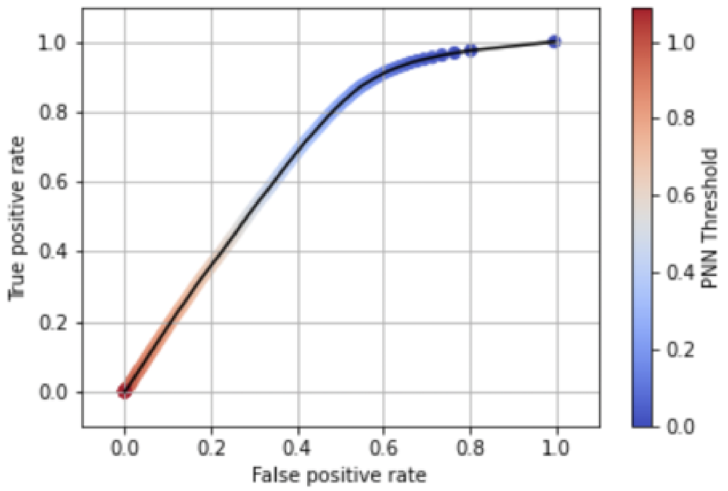


Test 8:
Template-Specific

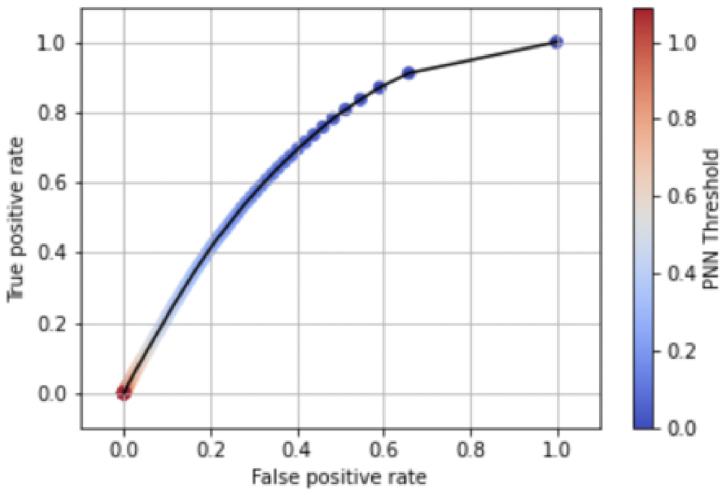
- Trained *without* Overlapping Data, Template-Invariant Criteria



Test 4: Raw Data



Test 5: Bandpass-Filtered



Test 6: Highpass-Filtered

Fine Tuning with Aftershocks

Considered a 'match': a template with one TP aftershock it detected

Tuned PNN models with overlap (raw showed improvement)

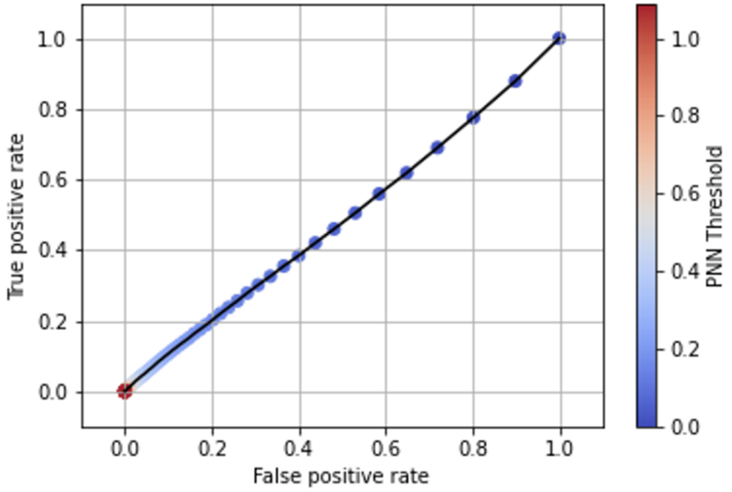
Test	TP	FP	TN	FN
1 (R)	606/21782	6738/169119	417899/255518	31247/10071

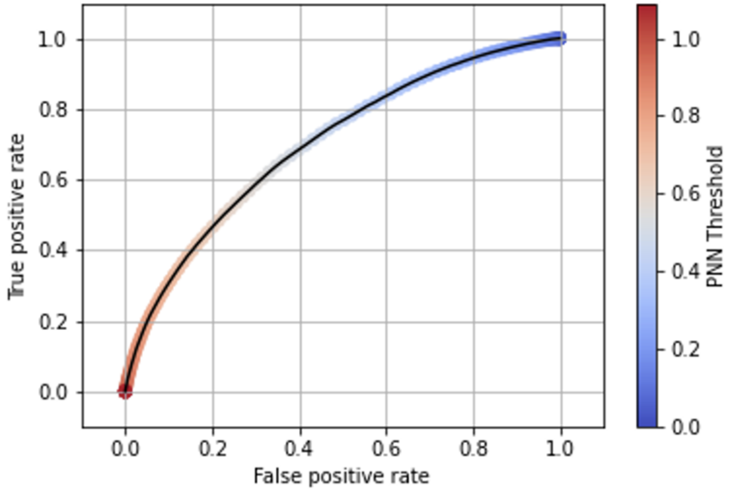
Test	TPR	FPR	TPR-to-FPR	AUC	Precision	F1
1 (R)	0.019/0.684	0.016/0.398	1.20/1.72	0.486/0.701	0.083/0.114	0.031/0.196

All numbers (except AUC) are for PNN threshold score = 0.5

Results are mixed and tuning process needs to be refined!

But some improvement observed…





Future Directions

- We need better metrics/criteria for how to define a match vs. a non-match
  - Explore template similarity & cross-correlation scores (for all waveform combinations)
- We need to think about model training
  - What's different between how original models were trained and why tuning with aftershock data improved it?
  - Geographic Distribution of data?
    - New validated 2011 Tohoku aftershock sequence data!
    - This region was better represented in the training dataset – so will the original PNN models work better here?
  - Original model use training data that were the same base waveform, but with different amounts of noise added.
    - Are those waveforms not different enough?
- We used a contrastive loss function, but would other loss functions be more appropriate? (e.g. Triplet Loss, as in Dickey et al., 2019)