Number of pages 33 (incl. appendices)
Number of appendices 3

TNO report

Identification and localization of micro-seismic events using the cross-correlation technique for the Ketzin CO_2 storage site

Date 05-08-2016

Author(s)
Maarten Reuver

Contents

1 Abstract

The Ketzin CO_2 storage site is a site that is intensively monitored for nearly a decade. An array, containing 3-C geophones and hydrophones, is located near the surface to monitor the pressure effects of the insertion of CO_2 storage. This array has been passively monitored from 2009 onwards. The sandstone saline aquifer located at 650 m depth (Förster et al., 2009) below the K2 high velocity layer proved to be ideal for CO_2 storage. The storage of CO_2 causes pressure build up and release in the subsurface (Park et al., 2013, Paap et al., 2014). This build up and release results in seismic events. In this study cross-correlation is used to identify these induced events using a waveform template from a pre-determined event. This master event occurred in 2010, originated from the storage aquifer depth. The cross-correlation method identified two extra events in the 2010 data set. These two events were found using the hydrophone receivers which were used because they have a higher signal to noise ratio (SNR) than the 3-C geophones. Two methods were tested to locate these events using; 1. the time difference between P-and S-wave arrivals, 2. the time difference between the P-wave arrival and the earthquake onset time (t_0). Method 2 proved to be most successful in locating the events. The main event and the two extra events found originate all three from the same location area, to the north east of the array. Because cross-correlation searches for the same waveform, events that are near identical can only be found. Events originating from a different location will not be detected. More events can still be found using the data sets of the remaining years. However, researching the 2010 data set revealed that even though events are found, they are of small magnitude and small in number.

2 Introduction

The storage of CO_2 in saline aquifers and empty gas fields has been of industrial interest for more than a decade. Storing CO_2 can make fossil fuel burning a climate neutral process by capturing the gasses produced by industrial processes. There are many offshore projects of CO_2 storage but only few onshore. The amount of CO_2 that can be injected differs per field and can be up to many megatons. However, injecting these large amounts of gas into the subsurface might result in seismicity due to the pressure build up and release the injection causes (Martens et al., 2012). The effects of these pressure changes and if the gas migration can be monitored, has been investigated for the storage site of interest Paap et al. (2014).

The Ketzin storage site is an onshore saline aquifer just west of Berlin. This storage site is used as a storage and monitoring site and has a dense stationary seismic array located above it. The injection of the CO_2 gas started from July 2008 up until August 2013. The continuous passive measurements started in August 2009 and are still ongoing. Santonico et al. (2012) developed a detection algorithm to automatically detect seismic events of low magnitudes by quality control and suppressing the noise. This resulted in the detection of several thousand events per year of which over 99% is noise produced at the surface.

The aim of this study is to distinguish between actual seismic events caused by the pressure changes of the stored CO_2 (eg. events produced at depth) and false detections generated by noise. There are several techniques to find the real events. Widely used methods are; time domain analysis, frequency domain analysis and cross-correlation. Time domain analysis detects events by searching for large amplitude differences over a short amount of time (short term average STA) compared to a long amount of time (long term average LTA). This technique is relatively easy and is widely used to detect seismic events with large signal to noise ratio (SNR) as for example done by Freiberger (1963). Frequency domain analysis searches for a particular spike in a frequency band or a certain frequency combination. Seismic events consist of both a P- and S-wave. The P-and S-wave have a different frequency band which can be identified separately. For this technique to work properly you need an estimate of the frequency band you are looking for and the event must not be of similar frequency as the ambient noise. Cross-correlation is a technique that can be applied if the waveform of a seismic event is known. It can detect seismic events of very small magnitude by cross-correlating the data set with the waveform template (Gibbons and Ringdal, 2006, Gibbons et al., 2007). Cross-correlation can only be used to detect events originated at approximately the same location as the template event.

In this study cross-correlation is used since we have a waveform template from a known event and we only want to look for seismic events originated from approximately the same location, the storage aquifer. A seismic event was detected using time-frequency domain analysis (Paap and Steeghs, 2014). Subsequently the magnitudes of a selection of events was calculated in a range of M_L = [-2.2,-0.3] (Paap and Steeghs, 2016). This analysis only worked during the night when the ambient noise was minimum. Since the Ketzin array has multiple receivers close to one another the SNR can be amplified by stacking the data to create a clearer signal for cross-correlation.

Events detected by cross-correlation must be verified to come from the depth of interest (the CO_2 storage field). In this study two methods are used to locate the new found events: 1. the difference between P- and S-wave arrivals, 2. the difference in P-wave arrivals with earthquake onset time (t_0). These two methods are then tested for multiple velocity models each consisting of a different amount of layers to investigate the effect of the layer interfaces.

3 Site characteristics

The Ketzin storage site is located near the village of Ketzin just west of Berlin. This was the first European onshore CO_2 storage site where CO_2 is injected and monitored. From July 2008 until August 2013 67,000 tons of CO_2 was injected (Boulenger et al., 2014). The insertion rate and pressure changes up to 2011 are shown in figure 1 (Martens et al., 2012) though the CO_2 injection lasted two years longer than shown in this figure.

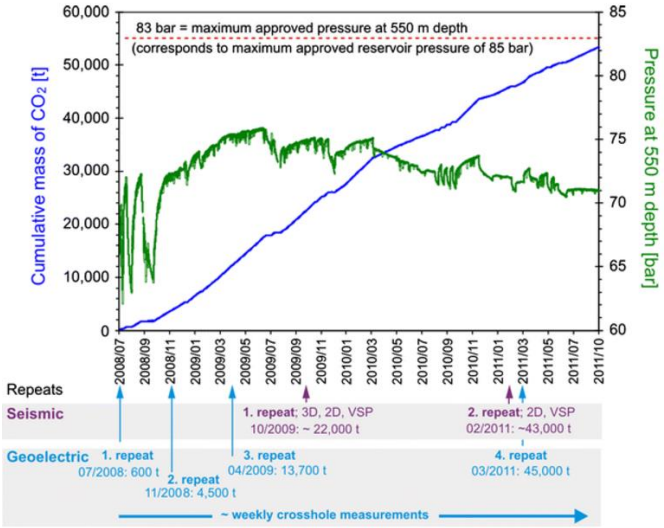


Figure 1
The blue line indicates the amount of CO_2 injected at the Ketzin storage site. The green line indicates the pressure measured at 550 m depth which is near the storage aquifer.
(Martens et al. 2012)

The array as shown in figure 2, consists of a multiple layered system. The total array is 120 m long. At the surface 7 3-C geophones are located. At 50 m depth 13 receivers consisting of both a 3-C geophone and a hydrophone are located with a spacing of 10 m. The hydrophones are only located in the buried positions since they need to be able to identify pressure caused waves. For a good vertical perspective, there are four 3-C geophones placed in the central borehole also consisting of this combination resulting in a total of six 3-C geophones positioned above one another with a spacing of 10 m. By looking at the arrival times of seismic events recorded by the vertical component from the central borehole receivers it can be determined whether an event came from the surface or from below the array. The array is oriented in approximately an East-West direction.

The subsurface of the site has been intensively researched by numerous techniques such as core analysis of the Ktzi201 well located at the site as shown in figure 3 (Förster et al., 2006) and 3D reflection seismics (Juhlin et al., 2007). The site lies on top of an anticline formed by a salt diaper 1500-2000m below the surface. The layer in which the CO_2 was injected is the Stüttgart formation. This is a saline aquifer consisting of sandstone and mudstone layers in total of approximately 80 m thick located with the base at 650 m below the surface (Förster et al., 2009). Just above this formation at a depth of 560 m lies a high density anhydrite layer with a P-wave velocity of approximately 5.3 km/s. Due to its high velocity this layer is also a sharp reflector for many incoming waves from near and afar.

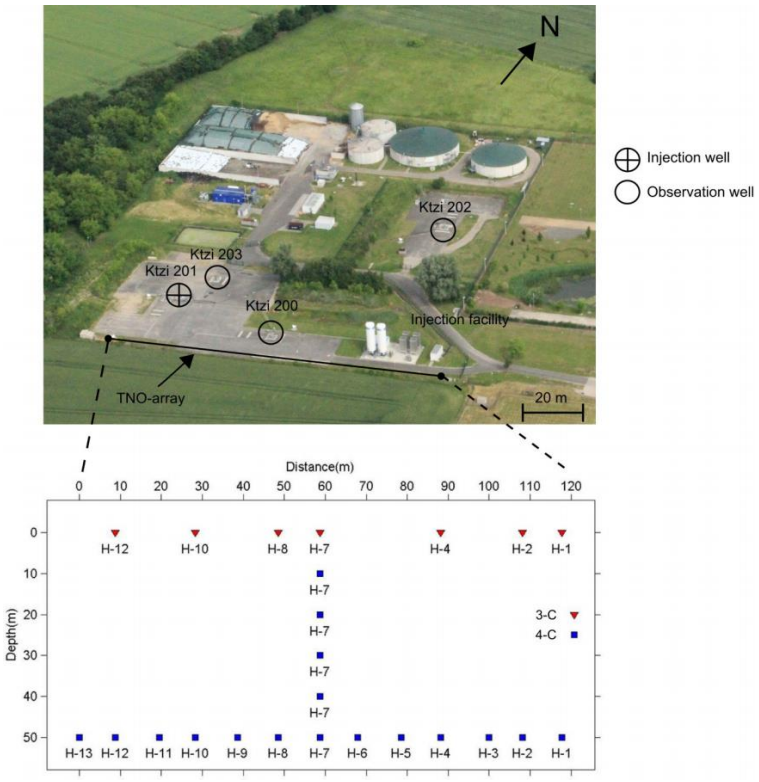


Figure 2

Top: An aerial view of the Ketzin storage site. Bottom: The 24 stations of the passive array. The surface station only have 3-C geophones. The buried station also contain hydrophones. (Paap et al. 2014)

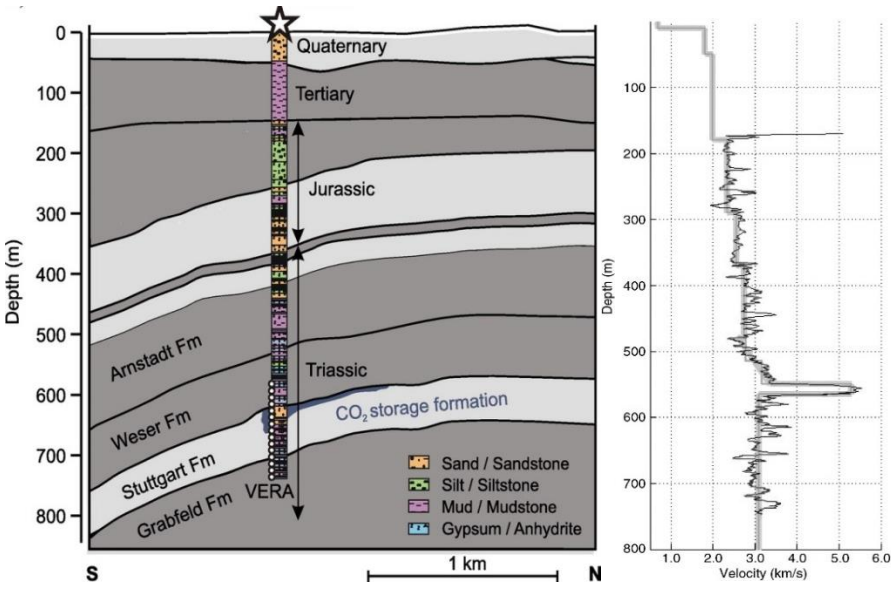


Figure 3

Left: The subsurface of the Ketzin storage site obtained from borehole Ktzi201. Right: The P-wave velocity profile obtained from borehole Ktzi201.
(Boullenger et al 2014)

4 Data characteristics

Seismic data was acquired continuously from July 2009 onwards. The sampling rate is 500 Hz. This results in a time step interval $dt = 0.002$ s. This dt holds for most of the time and most of the data. However, part of the data is sampled with 2000 Hz whenever there was a need to acquire more detailed data (e.g. active seismics or well drilling) which produces a dt of 0.0005. Continuous sampling at this rate results in a data set of more than 80 TB. Because of this size automated event detection is applied to this continuous data set. An event is detected if the SNR exceeds a certain threshold according to the STA/LTA (short term average/long term average) algorithm, i.e. by time domain analysis of the data set. The events are stored as data files with corresponding images of the traces from all 3-C geophones and hydrophones. The file length of each event is 4 seconds. For each year this results in 1600+ event files of only a couple of hundred kB each which is a massive data size reduction.

The site is located in a region with relatively low seismic activity. Large close by earthquakes are therefore unlikely to be measured except for the induced seismicity of the open air mining across the border in Poland. In the area several types of noise sources are active which will show up in the dataset such as the truck delivering the CO_2, the road nearby and agricultural machinery cause enough vibration to be detected by the array. All these sources yield false events. The expectation for the amount of actual real events from depth is that only a small number a year will take place, meaning less than a percent of the events that have a higher than average SNR.

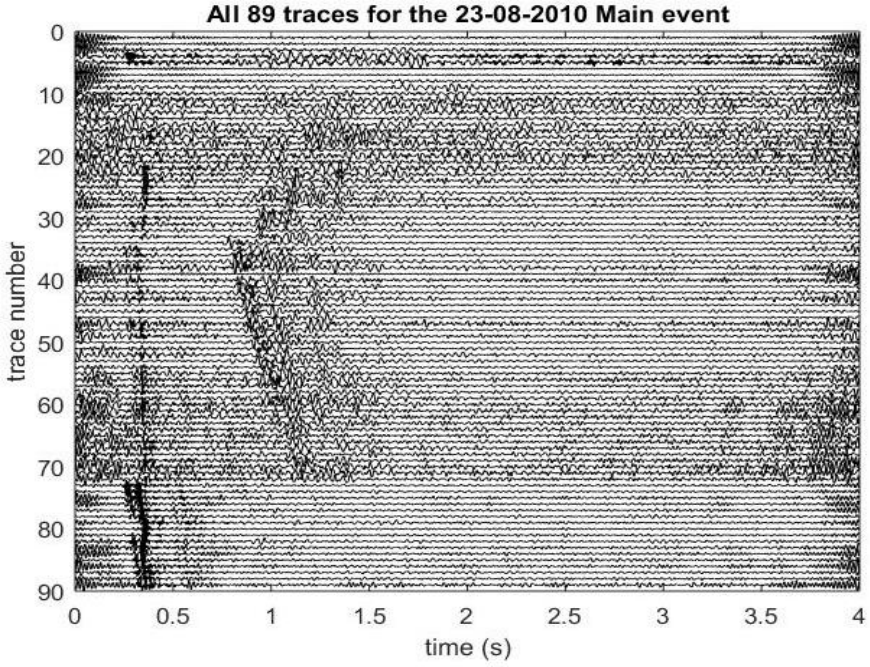


Figure 4

All 89 traces for the main event. The first 21 traces represent the 3-C geophones at the surface stations. Traces 22 up to 33 represent the 3-C geophones in the central borehole. Traces 34 up to72 represent the 3-Cgeophones at 50m depth. Traces 73 to 89 represent the hydrophone traces for both the central borehole as for 50 m depth. For a clearer view a bandpass filter has been applied (low 20 Hz, high 150 Hz).

5 Master event

5.1 A view of the master event

On the 23rd of August 2010 a micro-quake was detected with an STA/LTA algorithm. This event is taken as the master event with which all other data is compared. The traces for all the components are shown in figure 4. This figure shows all 3-C geophones and hydrophones for the stations shown in figure 2. The signal to noise ratio is not very high but the P- and S-wave can be distinguished very well. The hydrophones and X,Y,Z components of the geophones all show the waveform differently. The master event first has a sharp P-arrival and after a second followed by an S-arrival. It was estimated to be an event from approximately 600 meters depth which is the location of the CO_2 storage site. The event came from the North-East, since the wave front first arrives at the most North-East station as can be seen in figure 5 where trace 1 indicates station 1 in figure 2. This figure shows the Z component of the thirteen traces at 50 m depth.

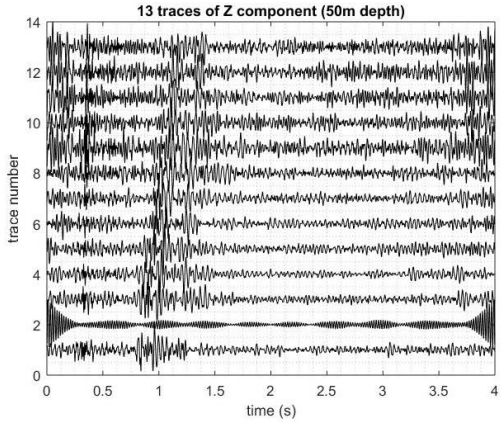


Figure 5
The 13 traces at 50 m depth for the Z-component of the Master event. A P-wave arrival can be seen at ~0.3 s and an S-wave at ~1.0 s. The NMO for the S-wave is much larger than for the P-wave. For a clearer view a bandpass filter has been applied (low 20 Hz, high 150 Hz).

The Z component indicates a clear P-wave and a less clear but still visible S-wave on all the traces except trace 2 since it is a defect trace. We look at the Z direction traces specifically because we are interested in a signal that originates from below the array. Signals originating from below the array can best be read on the Z-direction receivers. The 50 m depth receivers and the central borehole traces are able to identify this (figure 6). In the plotted data it is visible if the wave comes from below or above by viewing which arrival is first, 50 m depth (trace 0) or surface (trace 6). The surface receiver is highly distorted by noise.

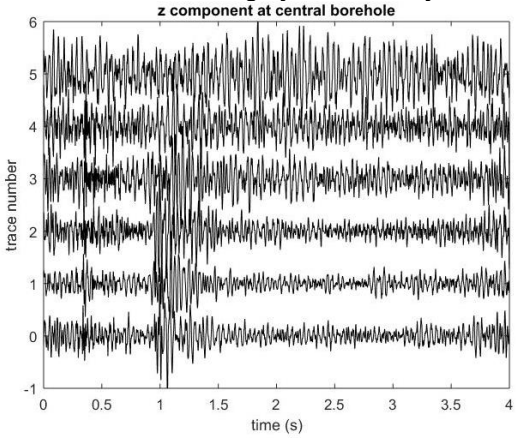


Figure 6
The central borehole traces for the Z-component. Trace 0 represents the trace at 50 m depth and trace 5 at the surface. Each subsequent trace shows the P- and S-wave arrival a bit later than the previous trace, meaning the event came from below 50 m depth. For a clearer view a bandpass filter has been applied (low 20 Hz, high 150 Hz).

The hydrophones also contain useful waveform data, but only for compressional waves. The hydrophone data for the master event is shown in figure 7. For the hydrophone the second trace does not malfunction opposed to the second trace for the Z-direction geophone. However, the eleventh and twelfth hydrophone trace are of opposite polarity and were flipped in order to compare them to the other traces and have the same polarity. The hydrophones will only show P-wave arrivals since it only shows pressure caused waves. They show a high SNR compared to the three components of the geophones.

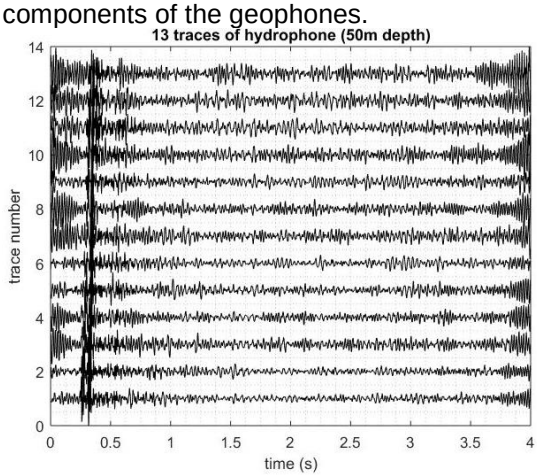


Figure 7

The 13 hydrophone traces at 50 m depth for the Master event. The hydrophone traces clearly only show the P-wave arrival at ~0.3 s, where for the 3-C geophones the S-wave could be seen at ~1.0 s. For a clearer view a bandpass filter has been applied (low 20 Hz, high

The X and Y (or North and East) components contain the clearest expression of S-waves and can be seen in figure 8. Since the geophones are not perfectly oriented in the North-South direction these components must first be corrected in order for all geophones to have the same orientation. For both components the difference is shown between the rotated and not rotated traces. It can clearly be seen that by rotating the traces the phase alignment correspond better with one another. The Z component did show an S-wave arrival. However, the S-wave is better distinguished at the X (East) and Y (North) components since the seismic wave from the 23-08-2010 master event comes from below the array. Therefore, the P-wave arrival is not clearly visible for these two components and should not be evaluated.

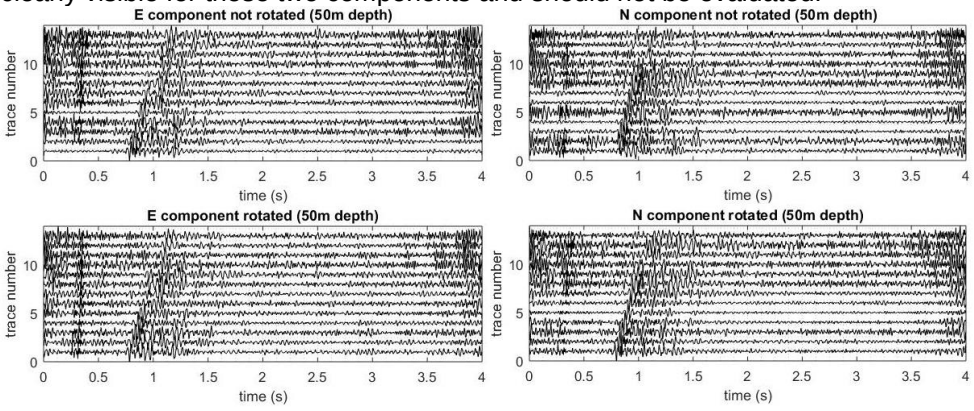


Figure 8

Left: The East component for the Master event. Right: The North component for the Master event. For both East and North, the top figure is not adjusted in rotation and the bottom figure is. The difference between the rotated and not rotated is that the rotated has a better alignment of the 13 traces. For a clearer view a bandpass filter has been applied (low 20 Hz, high 150 Hz).

5.2 Frequency filtering

For all components the traces in this figure are frequency filtered. The event was first distinguished by looking at the time frequency domain. The entire frequency band ranges from 0 to 250 Hz. For the master event this band is shown in figure 9. This figure shows that the lower frequencies, i.e. the frequencies below 20 Hz, have the highest amplitude values and dominate over the higher frequencies. This dominant high frequency band is noise generated. From the detected event the P-wave is present most on higher frequencies and the S-wave mainly around 50 Hz which is also a high amplitude value. In order to filter out the noise as much as possible the lower most frequencies are left out. The low bandpass frequency filter is set to 20 Hz for the entire data set. The highest frequencies should also be filtered out. They cause the data to be too spiky and may distort the waveform too much. The high bandpass frequency is tested for a variety of frequencies to see which one gave the best wave form. The parameters used are shown in table 1. For the master event, the 150 Hz is taken as the high bandpass frequency. The effect this bandpass has on the frequency spectrums is shown in figure 9 and the effect it has on the waveform is shown in figure 10. Figure 10 shows that without the dominant lowest 20 Hz the noise is suppressed greatly and the waveform is much clearer visible. The waveform was inspected for all high bandpasses stated in table 1. This revealed that a bandpass of 150 Hz showed the waveform best. The time-frequency domains for the unfiltered and filtered data are plotted for comparison in figure 11. When the high bandpass is filtered at a lower frequency the trace signal becomes smoother and with higher values more spiky. When searching for other events, the 150 Hz frequency band is also used since the frequency band of other events must be filtered identically.

Table 1

Parameter	Values
Low bandpass frequency (Hz)	20
High bandpass frequency (Hz)	130/150/200
Sampling rate frequency (Hz)	500 or 2000
dt (s)	0.002-0.0005

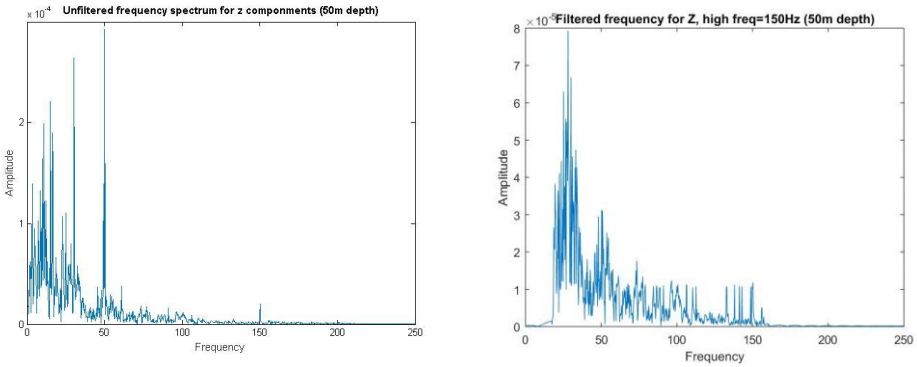


Figure 9

Left: The full frequency spectrum from the master event. Right: The filtered frequency spectrum for the master event at a low frequency band of 20 Hz and a high band of 150 Hz. Both figures are from trace 1 at 50 m depth.

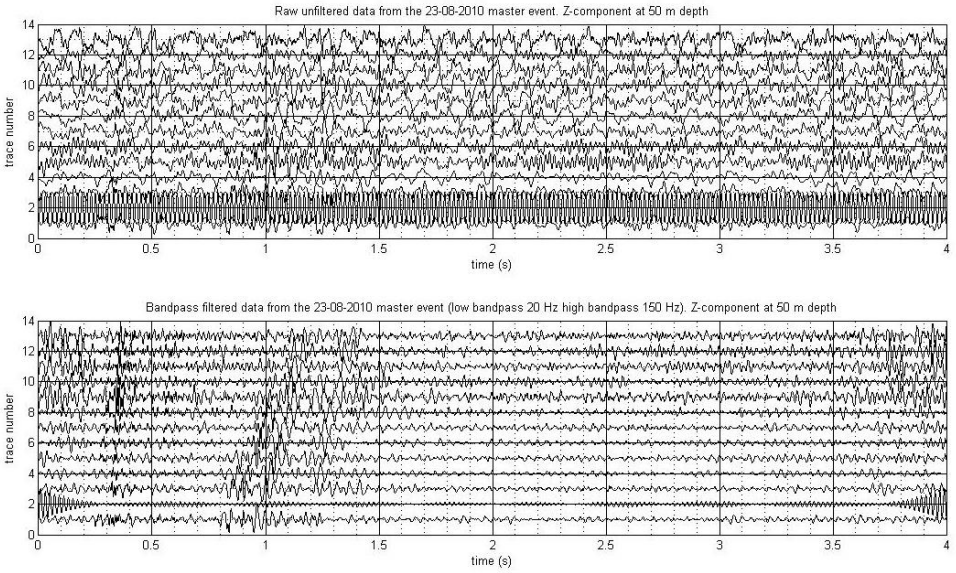


Figure 10

Top: Unfiltered data from the master event for the 13 Z-component traces at 50 m depth. Bottom: Filtered data for the same traces at the same time. The low bandpass is set at 20 Hz and the high bandpass at 150 Hz. Filtering the data with these bandpasses has a large effect on the waveform. In the bottom figure (filtered data) the event is much better visible. Leaving out these frequencies filters a lot of noise and leaves a cleaner signal.

5.3 Stacking

Stacking the master event data is essential to enhance the signal to noise ratio. The master event already has a relatively good ratio compared to the continues data. Before stacking the data a NMO (normal move out) correction for each trace must be made. The NMO of the P-wave is smaller than that of the S-wave by a factor of 5. By summing the traces of one event that have a similar waveform, the amplitude value of the P- and S-wave arrivals will increase much more over the noise. The noise, being random, will most likely cancel out because the noise patterns per trace are not coherent and when stacking 12 or 13 traces the positive cancels the negative amplitudes. Finally the stacked trace can be used as a waveform template to compare to other events. The stacked trace is an example of how much the SNR increases and is illustrated in figure 12 for the Z-component and hydrophones combined with all other traces. The stacked trace is normalized again. For the Z component the S-wave mostly cancels out since the NMO for the P-wave is different than for the S-wave (figure 12). Since the hydrophones only show pressure caused waves a stack is performed for the P-wave. The SNR for the hydrophones is largest of all and produces the clearest signal. The S-wave can be stacked on the X (East) and Y (North) component since it is most dominant on these traces and will therefore produce the highest SNR.

Not all traces for each receiver, either 3-C geophone or hydrophone, work continuously. Some traces are defect for certain periods of time. This must be taken into consideration when inspecting the data since they often highly distort the signals. This defect can hold for the all components.

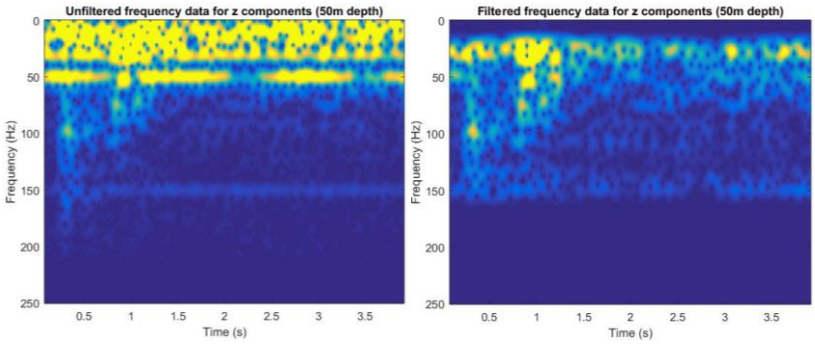


Figure 11

The time frequency spectrum for the master event. Left: the full unfiltered spectrum. Right: the filtered spectrum at a low frequency band of 20 Hz and a high band of 150 Hz. Both figures are from trace 1 at 50 m depth

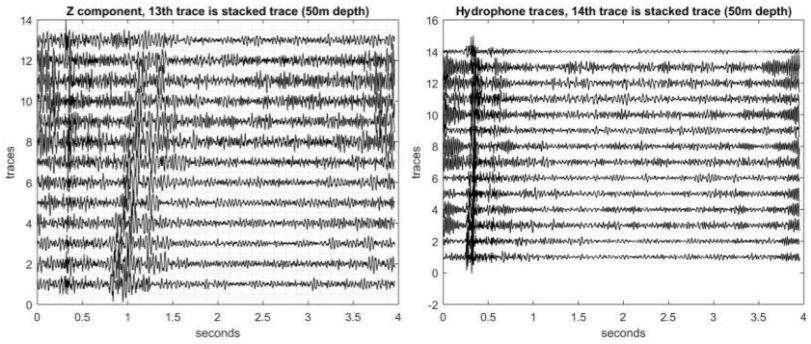


Figure 12

Left: The 12 active Z-component traces for the master event and on the 13th trace the P-wave stacked trace. Most of the signal flattens except for the P-wave which amplifies. Right: The 13 hydrophone traces for the master event and on the 14th trace the P-wave stacked trace. For the hydrophone the P-wave signal becomes even clearer than for the Z-component. All other noise disappears.

6 Methodology

6.1 Cross-correlation

The master event can now be used to detect other events with a similar wave form. To identify these events cross-correlation of the stacked master waveform is used with the full data set. Cross-correlation describes the resemblance between two traces a and b of length M and N.

$$C_k = \frac{1}{N+M+1}\sum a_p b_{k+p} \qquad \textit{(equation 1)}$$

Here k represents the lag. This lag can be positive or negative. For this study Pearson cross correlation is used because we want to calculate the resemblance between two linear variables which are normally distributed. The trace of the master waveform shall be vector $\bar{a}$ and is of length M, starts at t_a which is any chosen time and has time steps Δt which depend on the sampling rate (table 1).

$$a_{M,\Delta t}(t_a) = [a(t_a), t_a + \Delta t, .., a(t_a + (M-1)\Delta t)] \qquad \textit{(equation 2)}$$

The trace to be investigated is vector $\bar{b}$ and is of length N. It starts at time t_b and has time steps Δt and is given just as in equation 1. The way both vectors are described depends on the state of the lag. They are described differently if the lag is negative opposed to when it is positive. In appendix C a piece of the code is shown which describes both vectors as code 1. The Pearson correlation coefficient between these vectors results from equation 3.

$$r = \frac{\sum_{i=1}^{M}(a_i-\bar{a})\sum_{i=1}^{N}(b_i-\bar{b})}{\sqrt{\sum_{i=1}^{M}(a_i-\bar{a})^2}\sqrt{\sum_{i=1}^{N}(b_i-\bar{b})^2}} \qquad \textit{(equation 3)}$$

The piece of code calculating this is added in appendix C as code 2. This coefficient has a value in the interval of [0,1] and represents the degree of similarity between the two traces. A value of 1 corresponds to a 100% correlation and 0 to no correlation at all. This range is used instead of the standard [-1,1] because we are not interested if the waveform is of opposite polarity, but only in similarity. M has the length of the P- or S-wave of the master trace and will always be of smaller size than N which is a trace of a full waveform. Because M is much smaller than N the correlation is performed in several steps since $\bar{a}$ must be correlated fully with $\bar{b}$. The amount of correlation steps per trace can be varied depending on the precision required. The steps of the correlation must overlap each other because otherwise the waveform overlap is not big enough. This is illustrated in figure 13. The first and last 25% of the correlation outcome can't be trusted since too few correlation points between the two traces are evaluated. This is shown in the final line of code 2. Therefore the amount of steps taken must at least be equal to the length of $\bar{b}$ divided by half of the length of $\bar{a}$ as shown in equation 4 which results in the number of steps taken as shown in equation 5. The number of steps is rounded down to the nearest integer.

$$\text{stepsize}_{\max} = \text{N}\frac{2}{\text{M}} \qquad \textit{(equation 4)}$$

$$nsteps = \frac{M}{stepsize} \qquad \textit{(equation 5)}$$

The step size can be varied for precision purposes. A smaller step size means higher precision but longer computing time since more loops are made over the same data set. In this study a step size of 25 corresponding to 0.05 s is used for optimal point coverage and computing time.

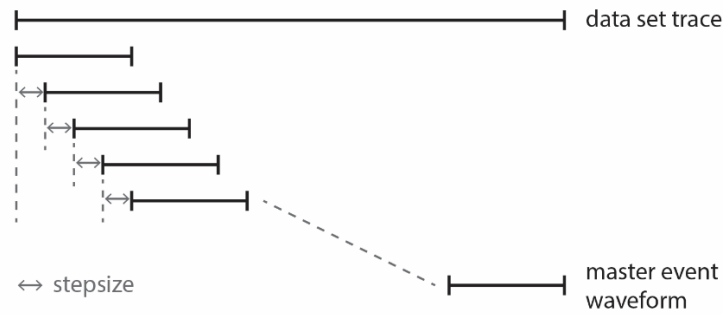


Figure 13
The correlation of the master waveform with the data set. The step size determines the precision of the correlation outcome.

Because of the stacked traces each component has a full waveform template with a high SNR available. The waveform template used differs in size depending on the component used. For the stacked Z and hydrophone trace only the P waveform is used which is a small part of the trace. For the X and Y component the S waveform can only be used.

The output of the code is a value for each file (each STA/LTA triggered piece of data) between 0 and 1. A fully coinciding waveform meaning a value of 1 for the correlation coefficient is only reached when correlation with the same trace (auto-correlation). The code output represents all the maximum correlation values for all traces from the investigated file averaged. Meaning if vector $\bar{a}$ represents a P-waveform in the Z direction at 50 m depth it will cross correlate with all the traces in the Z direction at 50 m depth each producing a maximum correlation value. The averaged output value per file is sorted form high to low. A threshold is set for the minimum value a file has to produce in order to be worthy of investigating. Based on insight gained from several test runs this value is set to 0.8 and results in approximately 30 files per year of which 1 or 2 are identified as actual micro-quakes. When time allowed, some files with a value just below 0.8 were also investigated.

6.2 Localization

Localization of the found events is essential to validate that they are events from depth. In particular, from the CO_2 storage area at approximately 600 m depth. To ensure a correct localization first a proper model needs to be build.

The subsurface of the Ketzin site is well known as previously discussed. The core analysis in figure 3 also shows the P-wave velocity with depth. The S-wave velocities can be calculated with a P- and S-wave relation according to Castagna et al. (1985) as stated in equation 6.

$$V_p = 1.16V_s + 1.36 \qquad \textit{(equation 6)}$$

The relation is for mud rock subsurface with velocities in km/s.

The localization is tested for two methods: 1. The time difference between the P- and S-wave arrivals, 2. The time difference between the P-arrival and t_0. These time differences need to be known for both the real events found as for synthetically produces events.

The synthetically produced events are built in a three dimensional point grid. The grid axis is oriented equal to the Ketzin array (North-East to South-West). Each point in this grid represents an event as a point source producing a synthetic seismic wave. The size of the grid differs per model run. The most used model has a size of x = [-1000 1000] and y = [-1000 1000] with a spacing of 200 m. Meaning an 11 by 11 grid in the x,y-plane. For each of these 121 grid points the model has 19 points in the z-direction starting at 100 m depth up to a 1000 m depth with a spacing of 50 m. Ultimately this forms a grid of 2299 points each with its own synthetic arrival times. The grid size can be changed for compute time optimization.

Each synthetic seismic wave produces a number of ray paths. Each ray path travels under a different starting angle. The range of angles at which a ray is emitted can be varied for computing time purposes. The angle range of the ray paths used lies between -0.4 pi and 0.4 pi with steps of pi/1000 and are calculated according to equation 7.

$$X(p) = \int_{-50}^{z_p} \frac{dz}{(u^2(z) - p^2)^{\frac{1}{2}}} \qquad \textit{(equation 7)}$$

The integral runs from the depth of the grid point z_p to the 50 m depth stations. The matching travel times are calculated according to equation 8.

$$T(p) = \int_{-50}^{z_p} \frac{u^2(z)}{(u^2(z) - p^2)^{\frac{1}{2}}} dz \qquad \textit{(equation 8)}$$

The travel times of the synthetic models can be compared with handpicked P- and S-wave arrivals from the events of interest. The arrivals are handpicked visually for both amplitudes on the traces and the time-frequency spectrum. The arrival times used are the average between the two. The outcome of the comparison is illustrated as a grid that shows the misfit for each grid point. The misfit (L_2) is calculated according to equation 9.

$$L_2 = \left((t_p - t_s)_m - (t_p - t_s)_d\right)^2 \qquad \textit{(equation 9)}$$

Subscript m is for the model arrivals for the P- and S-wave and subscript d for the data investigated. A lower misfit means a higher probability that the event originated from that particular grid point. It also shows the grid point with the lowest misfit as a black star representing the hypocenter of the event and the middle of the array with a black triangle.

To implement the depth-velocity profile several models are build consisting of a different amount of Earth layers. The real Ketzin model consists of 10 layers with different velocities. Each layer transition has an effect on the ray paths since the velocity changes. Ray paths are refracted by an interface according to Snell's law. The model only uses horizontal layers. To investigate the effect of the layers, different models with a different amount of layers are used. The velocities and the depth of the layer interfaces are shown in table 2. If a model consists of fewer layers the average velocity for that depth is used. The Ketzin receivers of interest lie at 50 m depth. Therefore, all ray paths in the synthetic models are recorded at 50 m depth.

The relative position between the three events can be calculated by subtracting the P-and S-wave arrival times of one of the events from a reference event, e.g. one of the other events, as given in equation 10.

$$\left(t_p - t_s\right)_{ref} - \left(t_p - t_s\right)_{new} \qquad \textit{(equation 10)}$$

If the outcome is smaller than 0, the new event lies closer to the receiver position. If it is bigger than 0, it lies farther away and if it is equal to 0, they lie at the same distance from the receivers.

	Nr. of layers	Layer interface depths (m)	Velocities m/s
Model 1	1	50 - 1000	2749
Model 2	3	50-290	2129
		290-520	2665
		520-1000	3100
Model 3	6	50-60	1800
		60-180	2000
		180-290	2300
		290-375	2500
		375-520	2750
		520-1000	3103
Model 4	10	50-60	1800
		60-180	2000
		180-290	2300
		290-375	2500
		375-520	2750
		520-555	3200
		555-572	5300
		572-632	3100
		632-652	2675
		652-1000	3100

Table 2
The layer models used for localization. Each model has a different amount of layers with different velocities.

7 Results

7.1 The events found

The first amount of runs were made on the 2010 data set. In this data set two new additional real events were located using the stacked hydrophone P-waveform and a high band frequency filtering at 150 Hz. The two events took place on the 22nd of February (from now on event 1) and the 25th of June (from now on event 2). Both events are displayed with all 89 traces in figures 14 and 15. For both events a P- and S-wave arrival can be distinguished. The P-wave arrives at ~1.8 s for event 1 and for event 2 at 0.3 s. The S-wave is for both a bit more than 0.7 s later. The time gap between the P- and S-waves coincides very well with the main event. The frequency domain for both events and that of the main event are shown together in figure 16,17 and 18. The frequency domains coincide quite well at most domains with one another. Event 2 has a high amplitude at 50 Hz caused by noise but its second highest at 25 Hz. The higher frequencies, larger than 100 Hz, are present at all events. The main difference between the three events is the signal strength. Event 1 has the highest amplitude values. This is also visible in figure 14 since the SNR is relatively high compared to the other two (figure 4 and 15).

Both events originate from below the surface as can be seen in figure 19. Here the P-wave in the central borehole is displayed for the Z component. The wave arrives first on the 50 m depth receiver followed by the shallower receivers subsequently. Since the central borehole shows that the two events originate from below the receivers the location for both can now be determined.

7.2 Grid size and localization

Using a small starting grid with x = [-500 500], y = [-500 500] and z = [-100 -1000], with dx=dy=100 m and dz=50 m, the localization for the main event as for the two events found lie at the outer boundary. After several tests with different grid sizes the eventual grid size that displays the events correctly and still has an acceptable computation time, is a grid with x = [-1000 1000], y = [-1000 1000] and z = [-100 -1000], with dx=dy=200 m and dz=50 m. The z range starts from -100 m since we want to look for events below the 50 m depth stations.

The localization is shown in appendix A and B. Appendix A shows the three events for three different models using the difference between P- and S-wave arrivals. Appendix B shows the three events for three different models but for the difference between the P-wave arrival and t_0. The different models used are shown in table 2.

The similarity of localization for the three events can be seen by comparing both Appendices. In both appendices the misfits for all three events, for each model used form table 2, are located at a similar location. No matter the method or the model used, the three events always have the lowest misfit concentration at the same location.

The difference between both localization methods can be compared well by comparing both appendices. Method 1 (the difference between P- and S-wave arrivals) illustrates the localization for most models not as good as method 2 (the difference between P-wave arrival and t_0). Only for the one layer models method 1 actually shows some resemblance to method 2. For the three and six layered models the two methods look nothing alike anymore. The more layers there are added the less method 1 appears to work. Therefore, the localization for the model with all layers

is only illustrated for method 2 in figures 20. This model has a high resemblance to the models in appendix B with the concentration of lowest misfits at the same location.

The relative location of the three events is determined according to equation 6. The order in distance from the receivers is event 2 < main event < event 1. The average difference from the 13 receivers between P-and S-wave arrivals for the three events are; event 2: 0.581 s, main event: 0.609 s, event 2: 0.664 s.

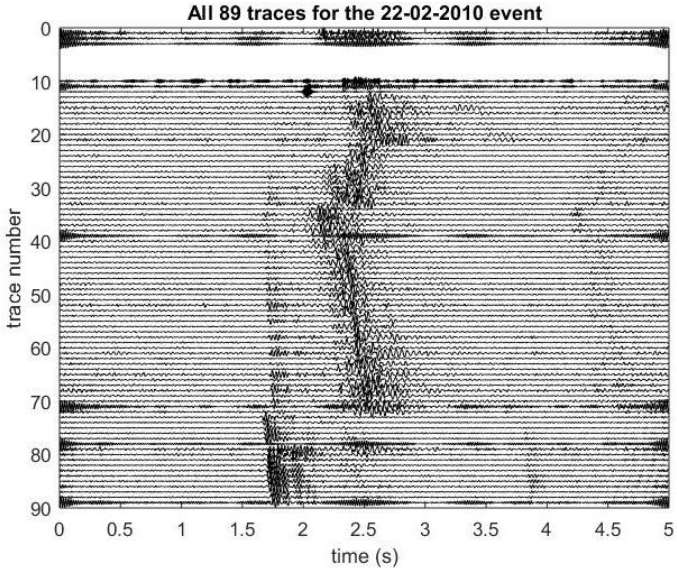


Figure 14

All 89 traces for the 22-02-2010 event (event 1). The traces are illustrated just as in figure 4. The P-wave arrival can be seen at ~1.75 s and the S-wave around 2.5 s. The S-wave is no visible for the last 17 traces which are the hydrophone traces.

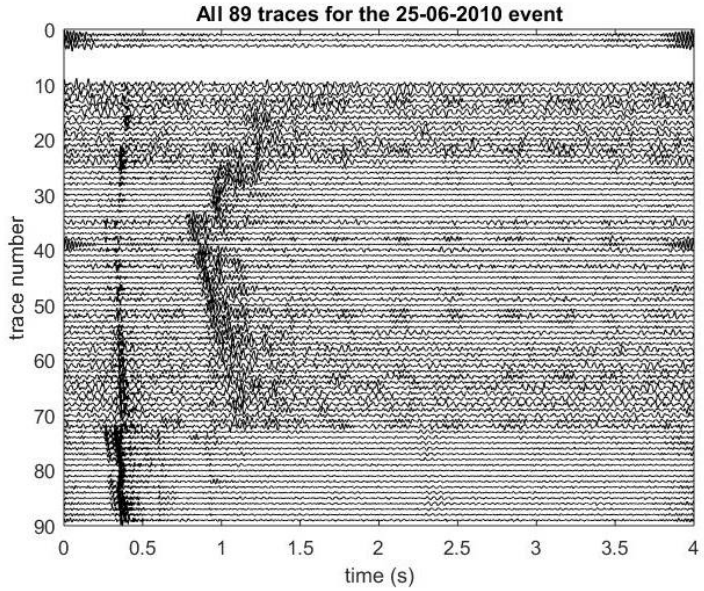


Figure 15

All 89 traces for the 25-06-2010 event (event 2). The traces are illustrated just as in figure 4. The P-wave arrival can be seen at ~0.4 s and the S-wave around 1.0 s. The S-wave is no visible for the last 17 traces which are the hydrophone traces.

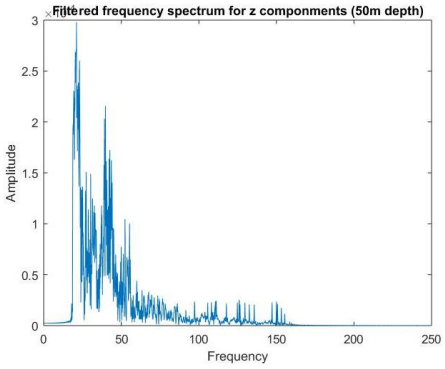


Figure 16

The frequency domain for the master event. The frequency is filtered at a low bandpass of 20 Hz and a high band pass of 150 Hz. Two peaks can be seen at 25 and 50 Hz which are mainly noise generated.

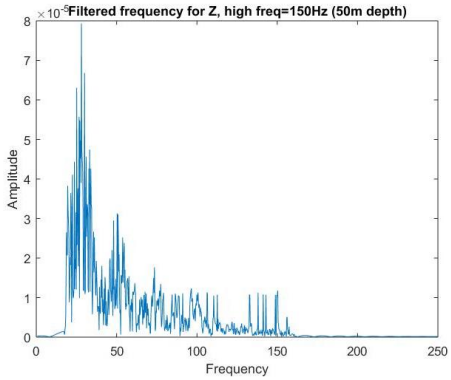


Figure 17

The frequency domain for the 22-02-2010 event (event 1). The frequency is filtered at a low bandpass of 20 Hz and a high band pass of 150 Hz. A large peak can be seen at 50 Hz which is mainly noise generated. The rest of the domain is similar to the other two events. However, the amplitude value is much larger.

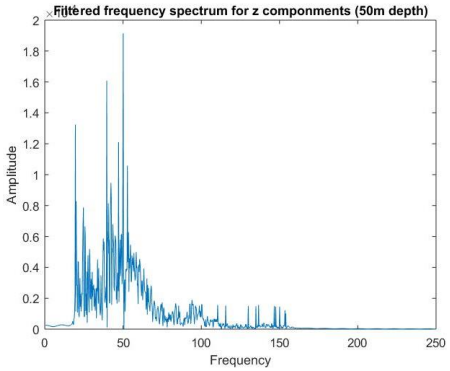


Figure 18

The frequency domain for the 25-06-2010 event (event 2). The frequency is filtered at a low bandpass of 20 Hz and a high band pass of 150 Hz. One main peak can be seen at 25 Hz which is mainly noise generated. The rest of the domain is similar to the other two events.

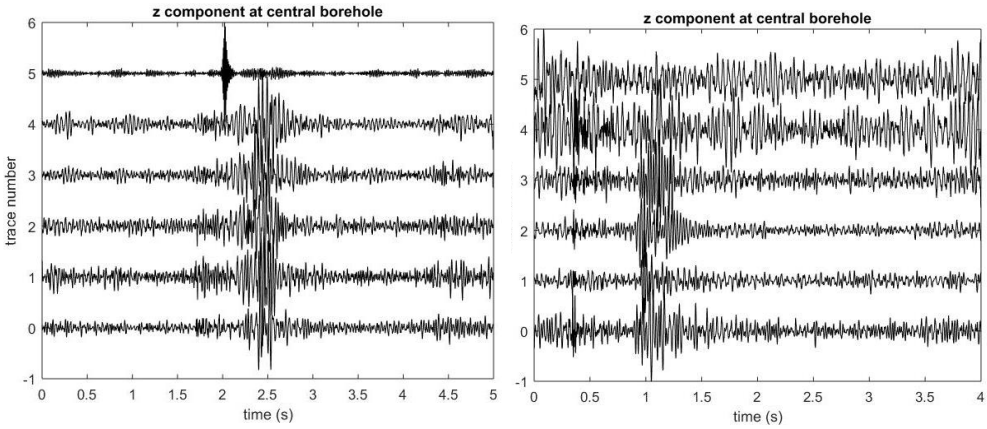


Figure 19

The central borehole traces for the Z-component. Trace 0 represents 50 m depth and trace 6 the surface with a spacing of 10 m between each trace. Left: For the 22-02-2010 event. The signal first arrives at the 50 m depth trace indicating an event from below. Right: For the 25-06-2010 event. The signal also arrives first at 50 m depth.

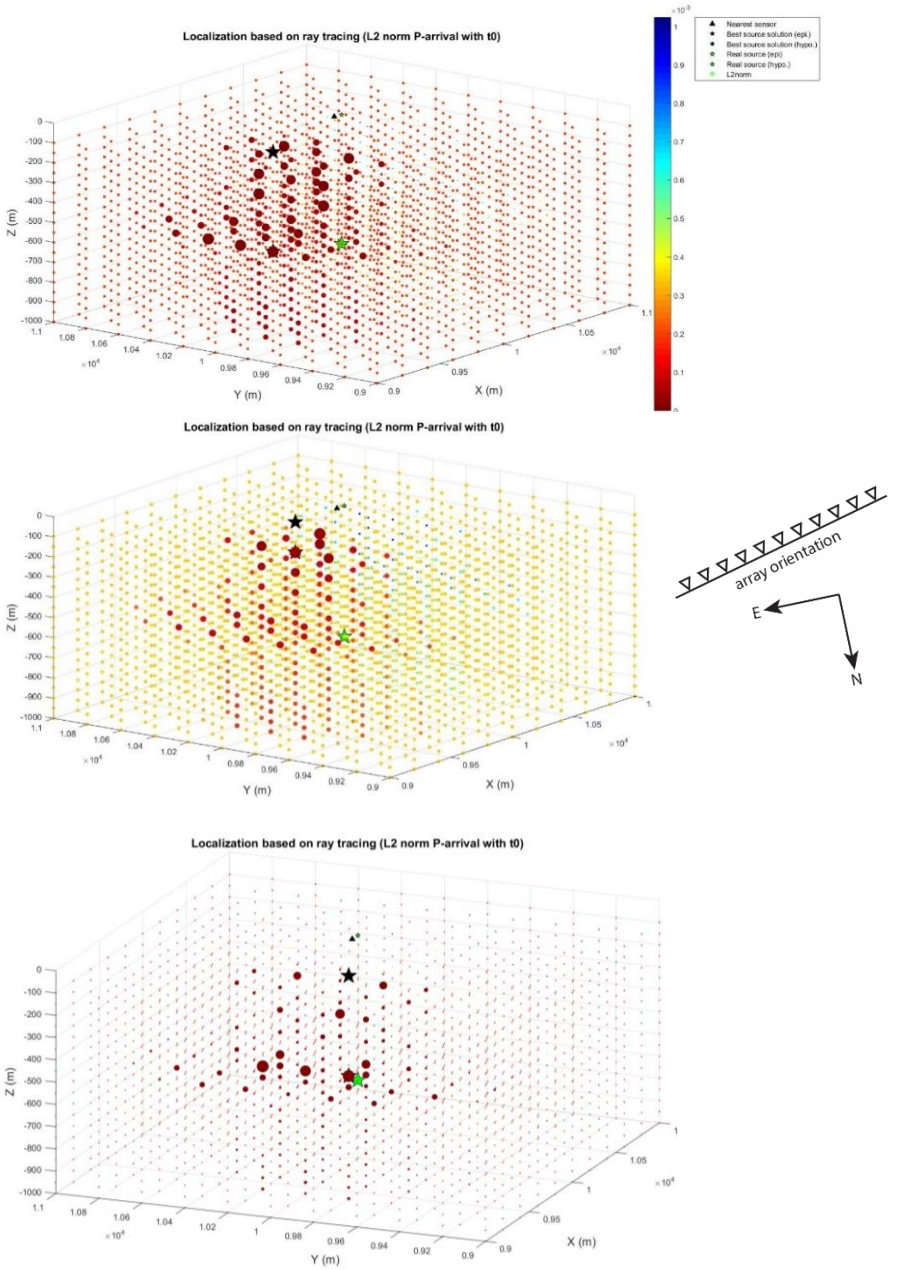


Figure 20

The localization of the three events from 2010 using the time difference between the p-wave arrival and the earthquake onset time (t_0). Model 4 from table 2 is used. Top: event 23-08-2010 (Master event). Middle: 22-02-2010 (event 1). Bottom: 25-06-2010 (event 2). The grid with size x = [-1000 1000], y = [-1000 1000] and z = [-100 -1000] with dx = dy = 200m and dz = 50m is illustrated. The black triangle at the surface indicates the center of the array. The array is positioned North-East to South-west. Looking at a figure means looking towards the South. Red color represents a low misfit, blue a high misfit. A larger circle indicates a lower misfit than a smaller circle. The black star at the surface indicates the epicenter of the lowest misfit and the one below the surface the hypocenter. The green star can be neglected.

For all three events, there are some low misfits near the surface. However, all three have the biggest concentration of large red circles located at 500-550 m depth, North-East. This indicates that the events came from this area. The master event and event 2 even have their lowest misfit in this area.

8 Discussion

8.1 Identification of the new events

Cross correlation is a technique to find the degree of coherence of two time sequences. The success of this technique, however, depends on many different parameters. It depends on the SNR for both the main event as for the data, the frequency band used, whether a combination or only one of the P- and S-wave form is used and if there are more than one master wave forms that can be used.

The SNR for this data set is very low. This complicates the use of cross-correlation because when the noise amplitude is similar as the event amplitude it will not be detected. Enhancement of the main waveform amplitude is therefore necessary. This worked very well by stacking the 13 traces at 50 m depth, especially for the hydrophones. The stacking produced a larger amplitude and the noise surrounding the signal got cleared as well. The amplitude of the main waveform template is enhanced and you obtain a clearer signal. The enhanced signal still has a higher SNR than most real earthquakes. Therefore, it will still correlate with some noise patterns. Due to this low SNR of the data the cross correlation outcome yielded many noise related events. Therefore, identifying the real events proved challenging.

The hydrophone signal proved to be most useful. Two events in 2010 were found using the hydrophone stack. The P-waveform from the main event for both the Z-direction as for the hydrophone waveform stacked very well and made a clear signal. The dataset contained much more noise for the traces in the Z-direction than for the hydrophone traces. This is probably why the Z-direction proved less useful than the hydrophone data. The two events found were determined to be from depth using the central borehole data (figure 19). Both events showed that the waveform first arrived at 50m depth and at the overlaying receivers subsequently.

So far only the 2010 data set has been investigated and only the P-wave has been used. There are 4 more years of data to be investigated and another 2 years of data coming in. Also, the stacked S-wave from either X or Y or a combination of a P- and S-wave as a correlation waveform template can be used to further specify the search. This will most likely result in more real events than those already found due to an extra correlation method. A combination of P- and S-waveform is, however, difficult because it cannot be stacked due to the different move out of both waves. The S-wave has a larger move out than the P-wave. A combination can be applied by first correlating with a P-waveform and if a high correlation value is found, the S-waveform is correlated over the following 1 second. This will filter some of the noise generated high correlation values which are found using only a P-waveform. This will save time researching the high correlation values.

The frequency band used while investigating this data is of high importance. Filter too few and the signal is too distorted. Filter too much and your signal loses its signature waveform. Therefore, it is important to first investigate the frequency band of the waveform. The frequencies lower than 20 Hz were filtered out to drain out a band containing much noise generated frequencies. From the three high bandpass frequencies, which were tested the waveform was best preserved with a high bandpass set at 150 Hz since the waveform amplitude was best preserved by not filtering too low and also not too spiky by filtering too high. This also holds for the two events found. The frequency domains coincide quite well at most peaks with one another, meaning a similar frequency domain as can be seen in figure 16, 17 and 18.

The fact that event 1 has the highest amplitude values means it has a larger magnitude than the other two events.

8.2 Localization of the three events

The accuracy of the localization of the three events depends on many different aspects. Besides the data quality it depends highly on model parameters such as the depth-velocity model, the size of the grid and its step sizes, the method used and the accuracy of the hand-picked arrivals.

The amount of grid points in the synthetic model used, i.e. the density of the grid, determines the accuracy of the localization. The grid size is tested for multiple sizes to investigate the distance between the receivers and the event locations. The grid size of x = [-1000 1000], y = [-1000 1000] and z = [-100 -1000] with dx = dy = 200m and dz = 50m is the optimal grid size. It shows enough points to establish a proper location for all three events. Even though a smaller dx and dy would make the localization more precise, this spacing does not take too much compute time.

The depth-velocity profile used producing the synthetic travel times is very important. It has an effect on the ray paths produced and therefore on the wave arrivals. The P-wave velocity was determined quite accurately. However, the S-wave velocity cannot. This S-wave velocity was determined using equation 6. However, this is only an average for the type of soil underneath the site. The actual S-wave velocities might differ from this relation, especially for the higher velocity layers. Therefore, the S-wave arrivals are less trustworthy than the P-wave arrivals.

The synthetic model is a 1-D model built in layers. Therefore, the layers are horizontally oriented and the ray paths can be produces symmetrical to both sides. However, the subsurface of Ketzin is not. The layers of the real subsurface are deformed by an anticline at 1500-2000 m depth. Even though they are not deformed in a large angle, this still affects the transmission angle of the ray paths produced. The ray paths always travel through the layer interfaces.

The subsurface below the Ketzin array up to 1000m depth consists of 10 layers. The velocity gradually increases with depth up to 560 m where a high velocity layers, the k2 layer, is located. This layer has a large effect on ray paths, since the velocity contrast with its upper and lower layer is over 2000 m/s. A velocity difference this high refracts the incoming waves with a high angle and influences the synthetic travel times greatly. The effect of this high velocity layer and the other layer interfaces was tested for the localization. Using the 10 layer model and method 1 (difference between P- and S-wave arrivals) the localization did not work properly. To investigate if this depends on the number of layer interfaces, the synthetic models with fewer layers were build. Appendix A shows that the use of more layers affects method 1 greatly. The layer interfaces refract both the P- and S-wave. The synthetic model produces ray paths with a small spacing between each ray path subsequently and not a full wave front as in figure 21. Therefore, the P- and S-wave, which both refract under different angles, will arrive at the surface receiver points with different spacing. Not all points will receive a P- and S-wave which will cause difficulties in localization using method 1. The smaller the spacing between each ray path, the more likely method 1 would succeed. A full wave front would be ideal, however, if this spacing will become smaller, the compute time will rise accordingly. For all models a ray path spacing of pi/1000 is used. This could be adjusted to test the difference in localization quality. For method 2 (difference between P-wave arrival and t_0) we only have the P-wave arrival at the surface receiver points. Therefore, this complication does not occur for this method.

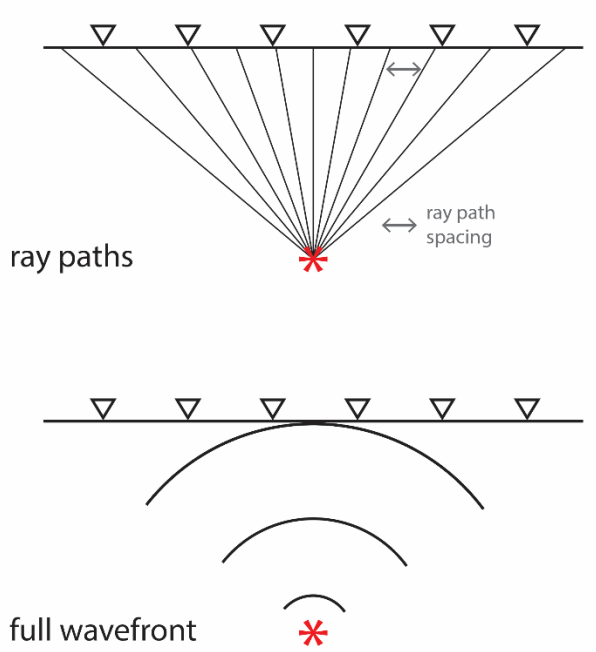


Figure 21
Top: Single ray paths produced from a source. Between each ray path there is a spacing determined by the angle at which it leaves. Bottom: A full wavefront produced from a source.
The top figure does not produce arrivals at each location at the surface but will have a spacing in between. Unlike the bottom figure with a full waveform.

The P- and S-wave arrivals for the three events are handpicked since the SNR is too low for an auto-picker. Because of the low SNR and because of human errors the handpicking of the events proved to be difficult especially for the S-wave arrival since the S-wave frequency coincides with a dominant noise frequency band which causes the S-wave signal to look like the noise generated signal. The P-wave arrivals were more clearly visible.

Based on the inaccuracy of hand picking arrivals in combination with the lack of a full wave front and the possible inaccuracy of the S-wave velocities, method 2 proved to be more successful in localizing the events than method 1 due to a simpler approach as can be seen by comparing appendix A and B.

Because of the previously mentioned complications; a very low SNR, inaccuracy by handpicking the arrival times and a synthetic model that cannot represent the real subsurface perfectly, the localization is not very precise. The location of the events have been determined to be in an overall area. All three are located in the same area at the same depth. Within this area it is determined by equation 6 that event 2 lies closest to the receivers. The main event and event 2 lie closest to each other because the difference between the P- and S-wave arrivals is are almost identical. Event 1 lies a bit further from the other two because the arrival time difference is a bit larger.

The cross correlation with this particular waveform, i.e. the main event waveform, resulted in very similar events with similar waveforms, originating from the same location and having a nearly similar frequency band. The CO_2 aquifer is several hundred meters wide. However, the cross correlation did not result in events from other locations. This technique combined with the quality of this data set only yields near identical events. They only differ in amplitude/magnitude shown in figure 16,17 and 18. The reason why the found events are approximately at the same location as the master event is because they were identified using correlation with the hydrophone component. This component measures in all directions at the same time. It does not have a fixed orientation like the 3-C geophones. It will correlate the master waveform only with waveforms from that particular direction. Correlating with three components (X,Y,Z) will likely also yield events from other directions since the incoming waveforms is measured in three located directions.

The amount of events that took place during 2010 might be higher than the three already found but at a different location. Although there might be more events at a different location and more in the other years to be researched, only a few amount of events take place each year per area and are small in magnitude.

9 Conclusion

The Ketzin CO_2 storage site is a widely researched area with a good quality seismic array. The injected CO_2 causes pressure build up and release that could cause micro-quakes of small magnitude. Over the past 7 years passive measurements have resulted in a large data set which contains a large amount of seismic information. The site has many background noise that is mostly human induced which makes it hard to identify the low magnitude micro-quakes. By finding an event originating from the location of insertion by STA/LTA analysis during the night, when noise is lowest, a waveform template could be made. With this waveform template other events from the same location could be identified using the Pearson cross-correlation technique. Even though the data has a very low SNR the correlation was successful and found 2 other events. These events were tested to be from depth and have a similar frequency domain as the main event. The best results were achieved with the hydrophone data since only pressure caused waves are recorded and contain the least noise of all traces. More events could have been found if the SNR was higher. The low SNR complicated the search for and localization of the events tremendously. If the entire data set for the extra 4 years is investigated and a correlation with the S-wave is performed, more events can still be found for this data set.

Waveform cross correlation searches for the same waveform in a data set. This has proven successful. However, it resulted in exactly the same kind of events which originated from the same location. The CO_2 storage aquifer is several hundred meters wide. Cross correlation for a SNR this low will not result in events originating from a location at the other side of the field. However, these pressure caused events might only originate at this side of the aquifer. Research of the 2010 data set revealed that a small amount of events per year, per particular area take place. Even though they do take place, they are of small magnitude and small in number.

Acknowledgement
This research was made possible by TNO. I would like to thank my supervisors Hanneke Paulssen, Bob Paap and Philippe Steeghs who all three helped me in their own helpful way. I would like to give special thanks to Bob Paap who provided me with some useful documents and codes.

10 References

Boullenger, B., Verdel, A., Paap, B., Thorbecke, J., & Draganov, D. (2014). Studying CO 2 storage with ambient-noise seismic interferometry: A combined numerical feasibility study and field-data example for Ketzin, Germany.*Geophysics*, *80*(1), Q1-Q13.

Castagna, J. P., Batzle, M. L., & Eastwood, R. L. (1985). Relationships between compressional-wave and shear-wave velocities in clastic silicate rocks.*Geophysics*, *50*(4), 571-581.

Förster, A., Giese, R., Juhlin, C., Norden, B., Springer, N., & CO2SINK Group. (2009). The Geology of the CO 2 SINK Site: From Regional Scale to Laboratory Scale. *Energy Procedia*, *1*(1), 2911-2918.

Freiberger, W. F. (1963). An approximate method in signal detection. *Quarterly of Applied Mathematics*, 373-378.

Gibbons, S. J., & Ringdal, F. (2006). The detection of low magnitude seismic events using array-based waveform correlation. *Geophysical Journal International*, *165*(1), 149-166.

Gibbons, S. J., Sørensen, M. B., Harris, D. B., & Ringdal, F. (2007). The detection and location of low magnitude earthquakes in northern Norway using multi-channel waveform correlation at regional distances. *Physics of the Earth and Planetary Interiors*, *160*(3), 285-309.

Juhlin, R. Giese, K. Zinck-Jørgensen, C. Cosma, H. Kazemeini, N. Juhojuntti, S. Lüth, B. Norden, A. Förster, (2007) 3D baseline seismics at Ketzin, Germany: the CO $_2$SINK project. Geophysics 72, B121-B132.

Martens, S., Kempka, T., Liebscher, A., Lüth, S., Möller, F., Myrttinen, A., ... & Kühn, M. (2012). Europe's longest-operating on-shore CO2 storage site at Ketzin, Germany: a progress report after three years of injection. *Environmental Earth Sciences*, *67*(2), 323-334.

Paap, B. F., and T. P. H. Steeghs. (2014) "Passive seismic monitoring at the ketzin CCS site-Magnitude estimation." *76th EAGE Conference and Exhibition 2014*.

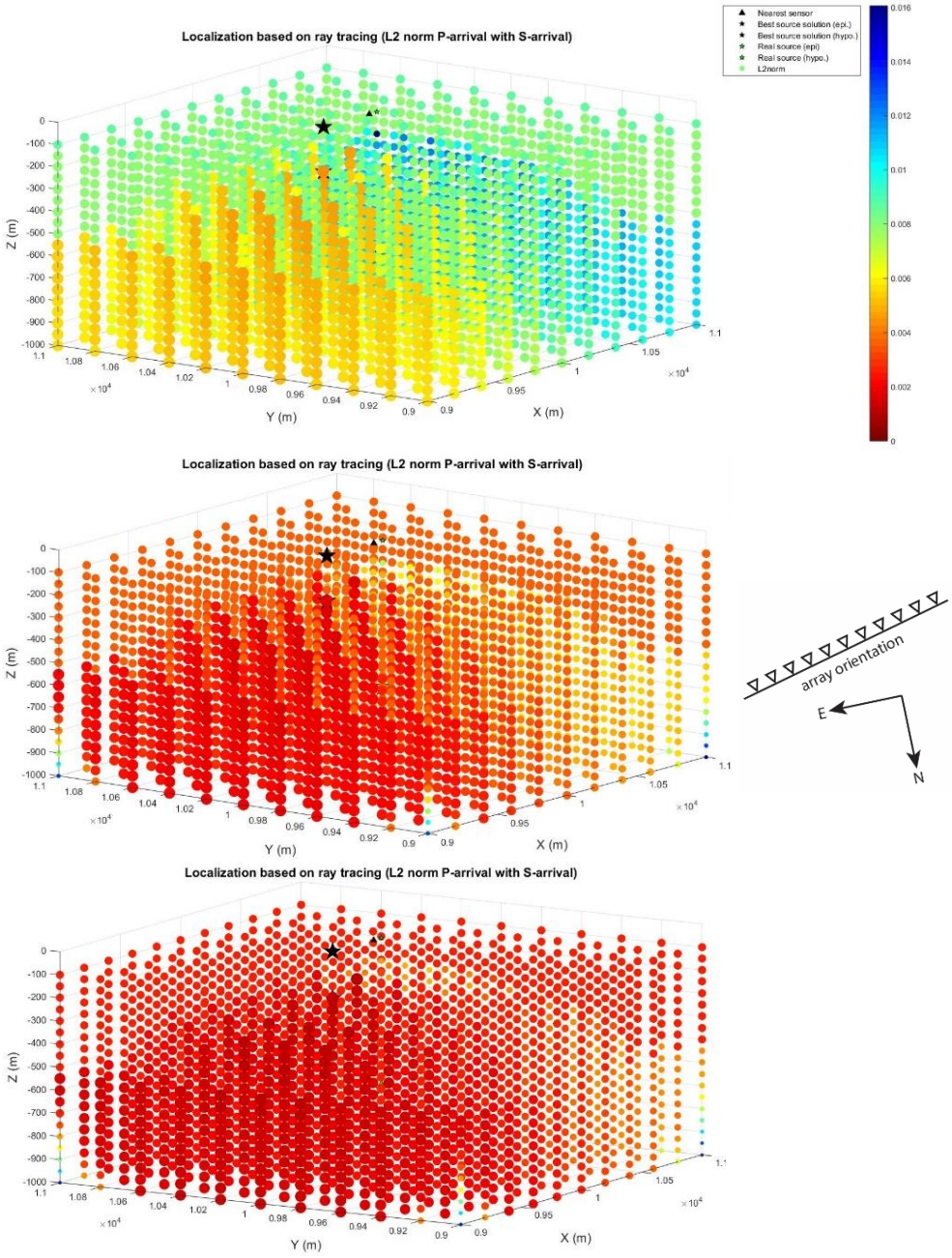
Paap, B., & Steeghs, P. (2016). Calibration of a local magnitude relationship for microseismic events using earthquake data. *Geophysics*, *81*(2), KS61-KS70.

Paap, B., Verdel, A., Meekes, S., Steeghs, P., Vandeweijer, V., & Neele, F. (2014). Four Years of Experience with a Permanent Seismic Monitoring Array at the Ketzin CO 2 Storage Pilot Site. *Energy Procedia*, *63*, 4043-4050.

Santonico, D., Zhang, X., Verdel, A. R., Meekes, J. A. C., & Arts, R. J. (2012, June). The first results of continuous passive surface seismic monitoring at the CO2 injection site of Ketzin. In *74th EAGE Conference and Exhibition incorporating EUROPEC 2012*.

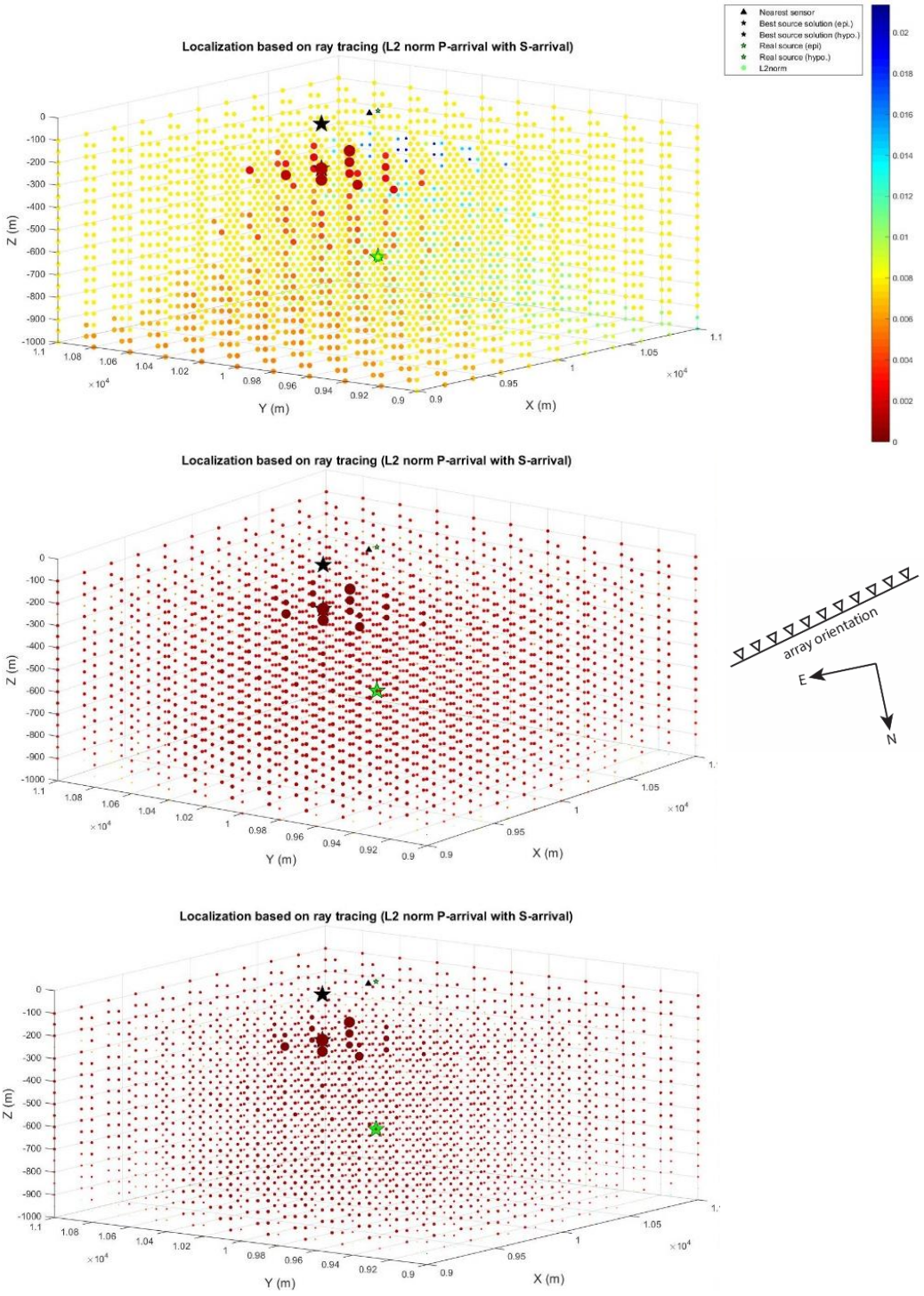
11 Appendix A

One layered model P-wave arrivals with S-wave arrivals



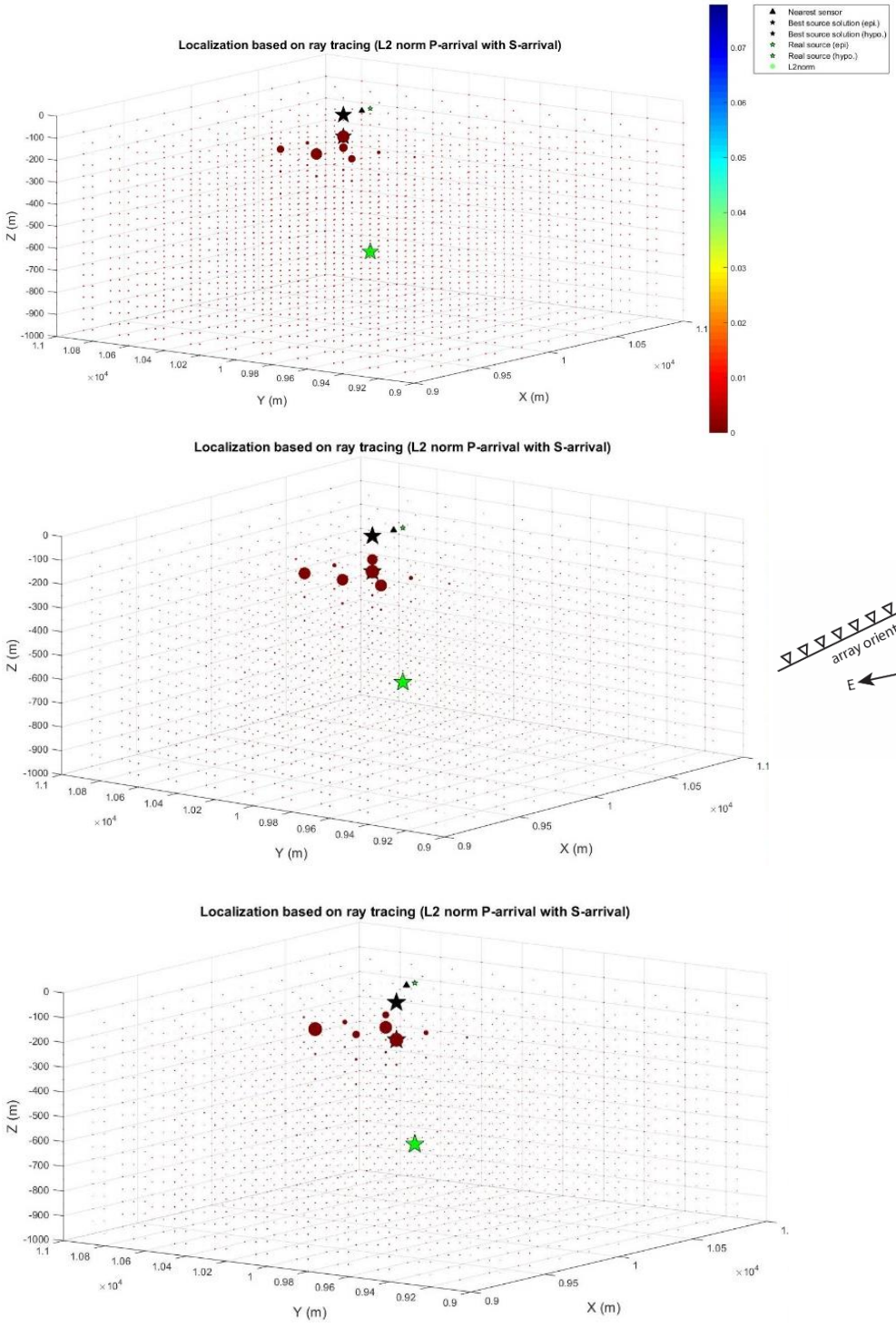
The localization of the three events from 2010 using the time difference between the p-wave arrival and the S-wave arrival. Model 1 from table 2 is used. Top: event 23-08-2010 (Master event). Middle: 22-02-2010 (event 1). Bottom: 25-06-2010 (event 2).
For all three events, the biggest concentration of large red circles is located at 500-600 m depth. This indicates that the events came from this area. However, it is a wide area and therefore not very precise. It can only be used to make an estimate of the location.
(see figure 19 for grid description).

Three layered model P-wave arrivals with S-wave arrivals



The localization of the three events from 2010 using the time difference between the p-wave arrival and the S-wave arrival. Model 2 from table 2 is used. Top: event 23-08-2010 (Master event). Middle: 22-02-2010 (event 1). Bottom: 25-06-2010 (event 2).
For all three events, there are some low misfits near the surface. There are no low misfits located below 200 m. This method (difference between P-wave arrival and S-wave arrival) does not work properly with this amount of layers.
(see figure 19 for grid description).

Six layered model P-wave arrivals with S-wave arrivals

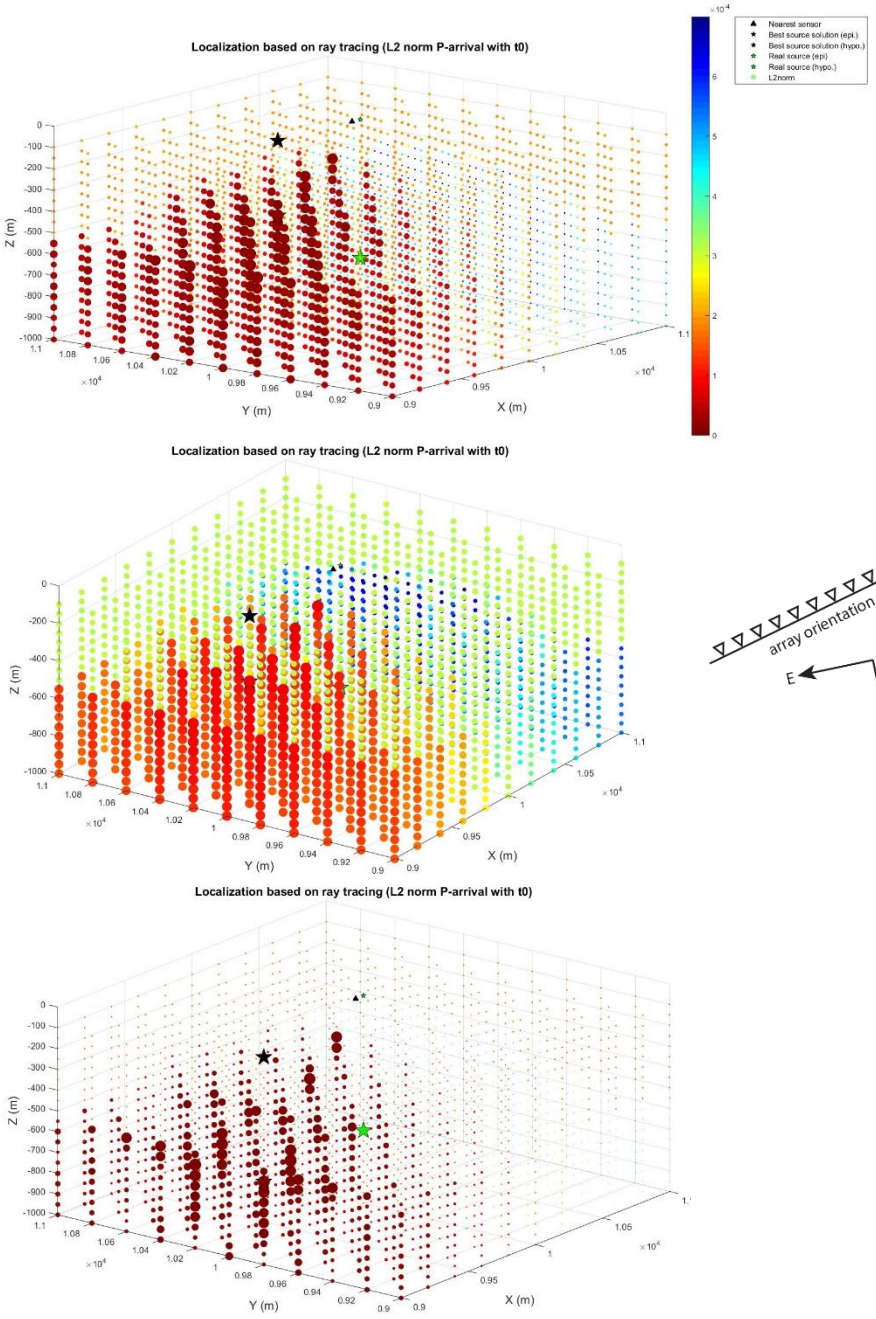


The localization of the three events from 2010 using the time difference between the p-wave arrival and the S-wave arrival. Model 3 from table 2 is used. Top: event 23-08-2010 (Master event). Middle: 22-02-2010 (event 1). Bottom: 25-06-2010 (event 2).

For all three events, there are some low misfits near the surface. There are no low misfits located below 150 m. All other points in the grid have a very high misfit, meaning this method (difference between P-wave arrival and S-wave arrival) does not work properly with this amount of layers. (see figure 19 for grid description).

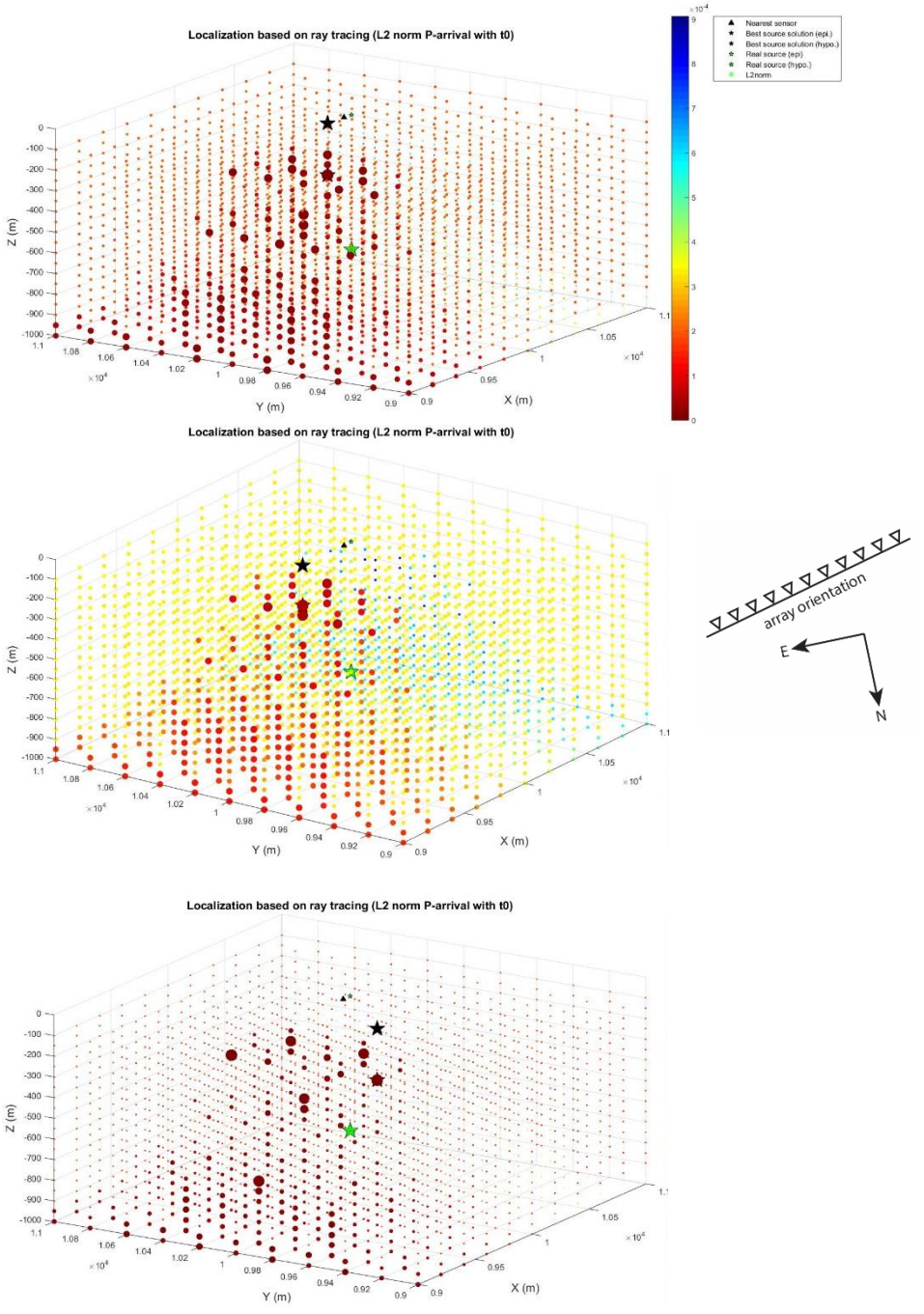
12 Appendix B

One layered model P-wave arrivals with t_0



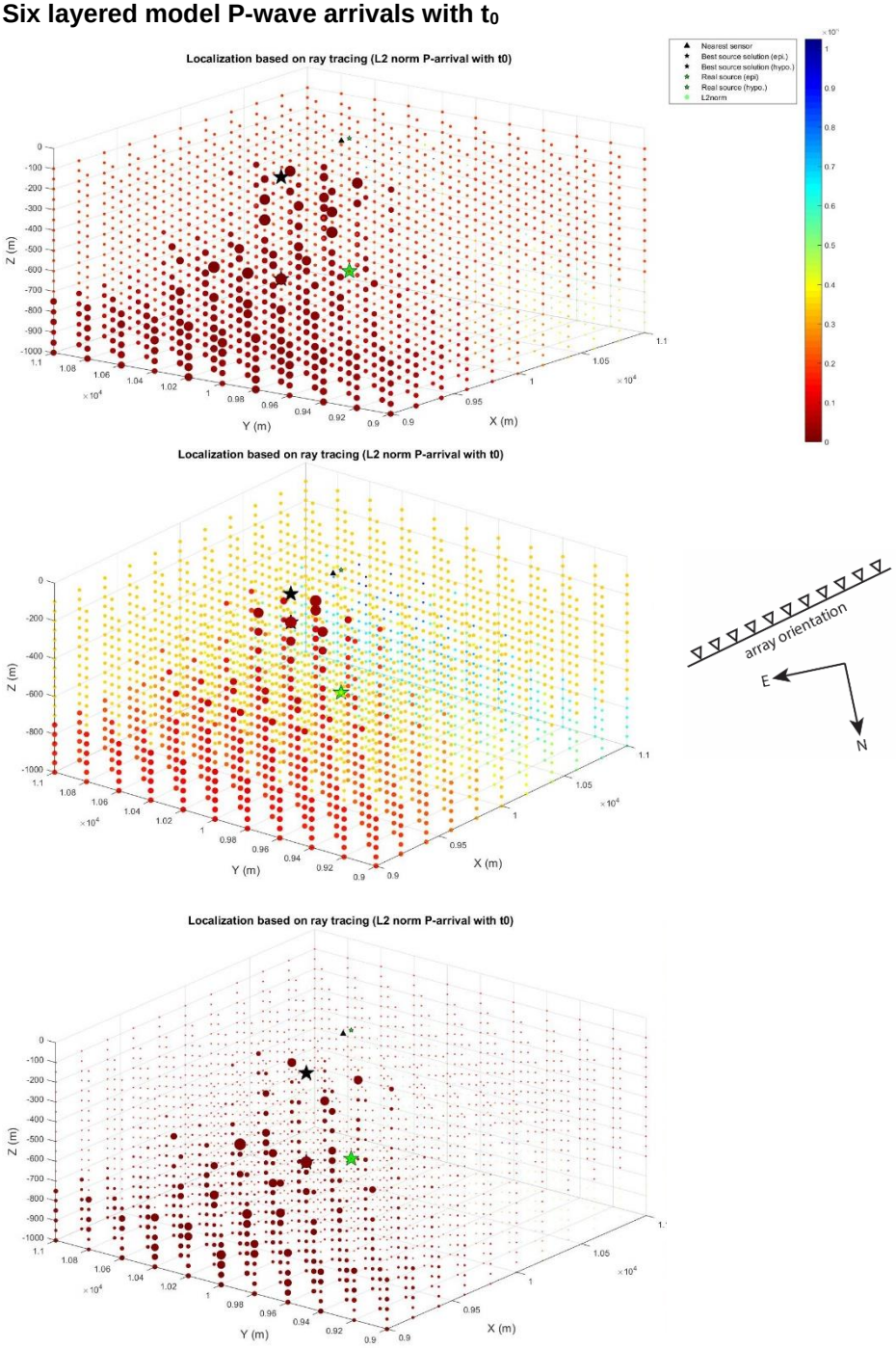
The localization of the three events from 2010 using the time difference between the p-wave arrival and the earthquake onset time (t_0). Model 1 from table 2 is used. Top: event 23-08-2010 (Master event). Middle: 22-02-2010 (event 1). Bottom: 25-06-2010 (event 2).
For all three events, the biggest concentration of large red circles is located at 500-600 m depth. This indicates that the events came from this area. However, it is a wide area and therefore not very precise. It can only be used to make an estimate of the location. These three localizations look very similar to the same model 1 figures in appendix A.
(see figure 19 for grid description).

Three layered model P-wave arrivals with t_0



The localization of the three events from 2010 using the time difference between the p-wave arrival and the earthquake onset time (t_0). Model 2 from table 2 is used. Top: event 23-08-2010 (Master event). Middle: 22-02-2010 (event 1). Bottom: 25-06-2010 (event 2).
The Master event and event 1 both show a concentration of low misfits at a depth of 600 m. For event 2 this is slightly less. However, all three figures have their lowest misfits at much shallower depth. The precision of this three layered model is the least of all models (table 1). The localization of these three figures are better than the same model 2 figures in appendix A.
(see figure 19 for grid description).

Six layered model P-wave arrivals with t_0



The localization of the three events from 2010 using the time difference between the p-wave arrival and the earthquake onset time (t_0). Model 3 from table 2 is used. Top: event 23-08-2010 (Master event). Middle: 22-02-2010 (event 1). Bottom: 25-06-2010 (event 2).
For all three events, the biggest concentration of large red circles is located at 650-700 m depth to the North-East. This indicates that the events came from this area. The area of misfits is smaller than for the other models. It is more precise and should be used prior to the other two. These three localizations nothing like same model 3 figures in appendix A and localize much better.
(see figure 19 for grid description).

13 Appendix C

```
for j=1:floor(nsteps)
    if  M-nstart(j) >= 100
            datacorr=b(nstart(j):nstart(j)+100);
            nshift = (length(datacorr)*2)-1;

            for i=1:nshift
                if i>n
                    j1 = 1;
                    k1 = (2*n)-i;
                    j2 = i-n+1;
                    k2 = n;
                else
                    j1 = n-i+1;
                    k1 = n;
                    j2 = 1;
                    k2 = i;
                end
          c5(i)=corcof_test(a(j1:k1),datacorr(j2:k2);
      else
      end

            maxc5(j)=max(c5([50:150]));
end
```

Code 1
C5 represents the correlation outcome between trace a and b. The start and end (for a: j1, k1 and for b: j2, k2) of trace a and b are described according to the state of the lag (k). With the last line the maximum correlation value is stored within the middle 50% of the correlation step.

```
function [cc]=corcof_test(a,b)

a=a(:,1);
b=b(:,1);

ma=mean(a,1);
mb=mean(b,1);

T=sum((a-ma).*(b-mb));

N = ((sum((a-ma).^2)).^0.5) * ((sum((b-mb).^2)).^0.5);

cc=T/N;
```

Code 2
The correlation function used in code 1. Trace a is the master waveform. Trace b is the data which is investigated.