# Fitting Statistical Models to DART® Buoy Data for Operational Tsunami Forecasting

Don Percival

Applied Physics Laboratory
Department of Statistics
University of Washington, Seattle

Joint work with Dancsi Percival (Google) and NOAA colleagues Don Denbo, Marie Eblé, Edison Gica, Paul Huang, Hal Mofjeld, Mick Spillane, Vasily Titov and Elena Tolkova

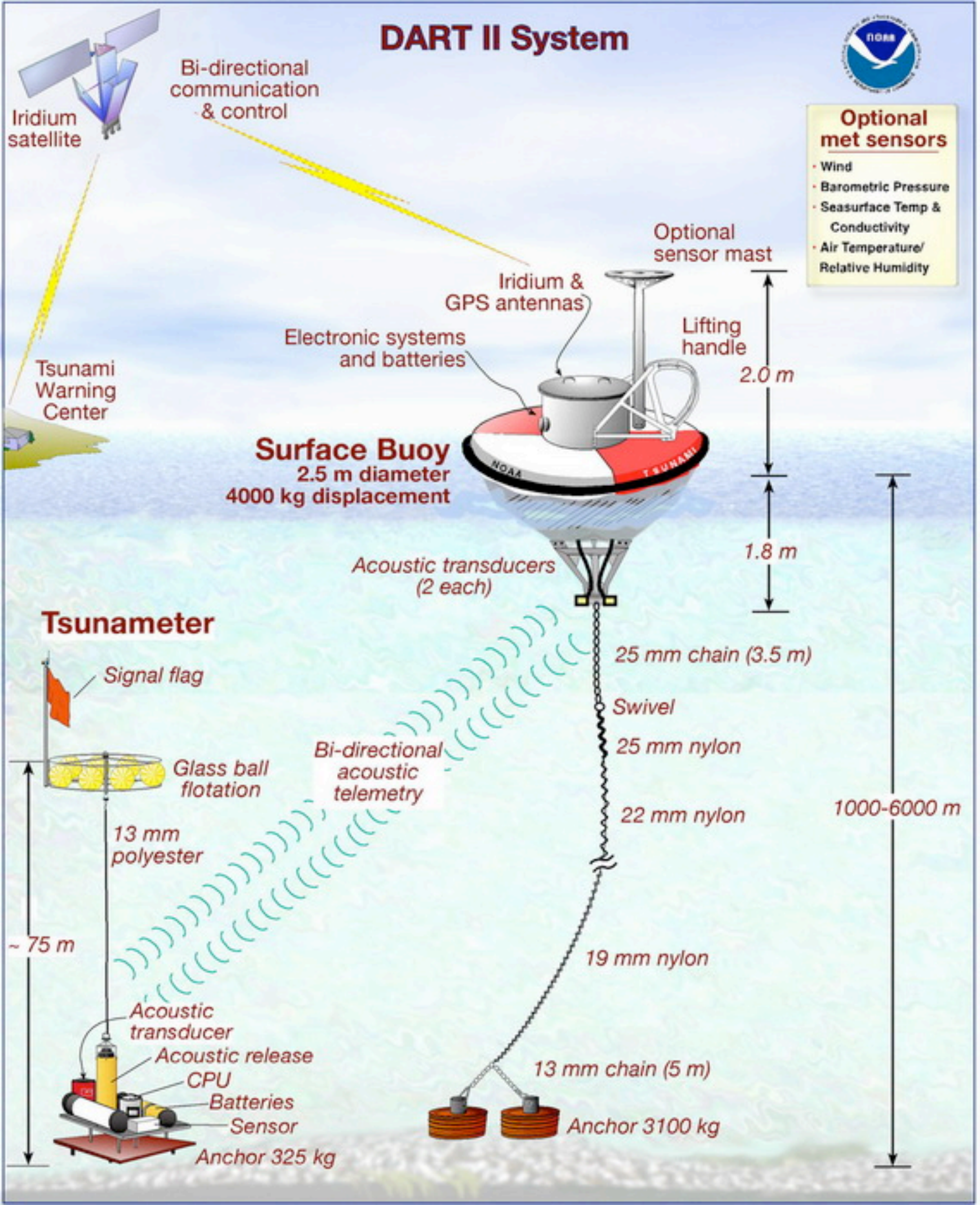
# Background: I

- destructive potential of earthquake-generated tsunamis well-known, with tragic recent examples being in the Indian Ocean (26 December 2004 – 230,000 estimated deaths) and off the coast of Japan (11 March 2011 – 15,893 confirmed deaths)
- due to rate at which tsunamis cross the ocean, possible to lessen effects on some coastal communities through advance warnings
- for US, warning centers in Alaska and Hawaii are responsible for issuing timely bulletins about impending tsunamis
- seismometers give first indication of a potential tsunami event
  - starting time of earthquake event
  - location of epicenter
  - initial estimate of magnitude

# Background: II

- seismic information alone is not enough to accurately forecast potential impact of tsunami
- led to development of Deep-ocean Assessment and Reporting of Tsunamis (DART®) buoys


RONALD H. BROWN
TSUNAMI


DART II System
Iridium satellite
Bi-directional communication & control
Optional met sensors
Wind
Barometric Pressure
Seasurface Temp & Conductivity
Air Temperature/ Relative Humidity
Optional sensor mast
Iridium & GPS antennas
Lifting handle
Electronic systems and batteries
2.0 m
Tsunami Warning Center
Surface Buoy
2.5 m diameter
4000 kg displacement
1.8 m
Acoustic transducers (2 each)
Tsunameter
25 mm chain (3.5 m)
Signal flag
Swivel
25 mm nylon
Bi-directional acoustic telemetry
Glass ball flotation
22 mm nylon
1000-6000 m
13 mm polyester
~ 75 m
19 mm nylon
Acoustic transducer
Acoustic release
CPU
13 mm chain (5 m)
Batteries
Sensor
Anchor 3100 kg
Anchor 325 kg

# Background: III

- prior to December 2004, six DART® buoys were deployed, all in Pacific Ocean
- as result of Indian Ocean event, US Congress mandated 33 more be deployed in Pacific Ocean and elsewhere (to date, eight other countries have also deployed DART® buoys)
- mindful of Percival's First Law

  'Real-time computer demonstrations are doomed to fail!'

  let's look at network of buoys and recent data by going to

  `http://www.ndbc.noaa.gov/dart.shtml`

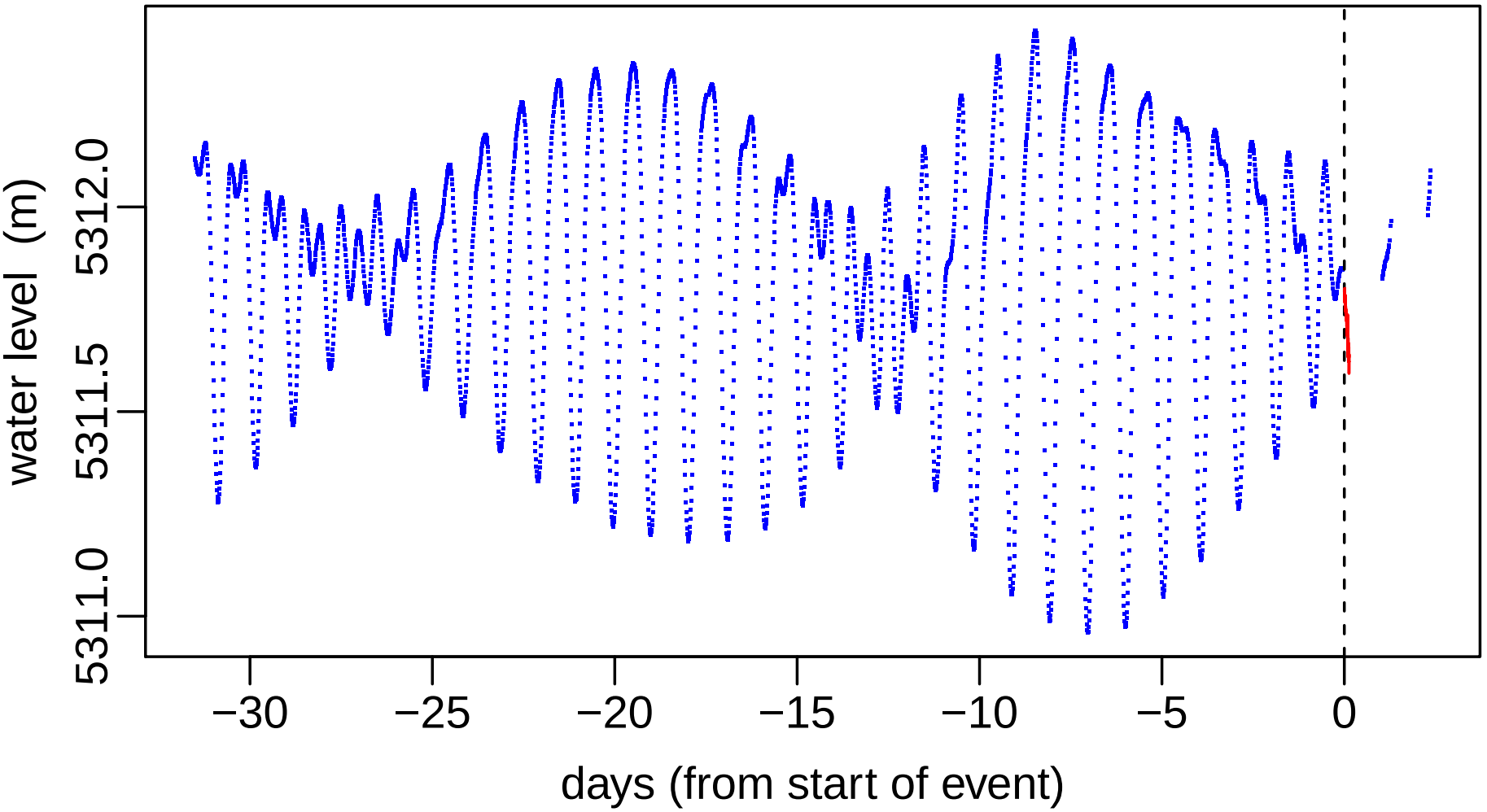
  and movie of Nov. 2006 Kuril Islands event downloadable from

  `http://nctr.pmel.noaa.gov/kuril20061115.html`

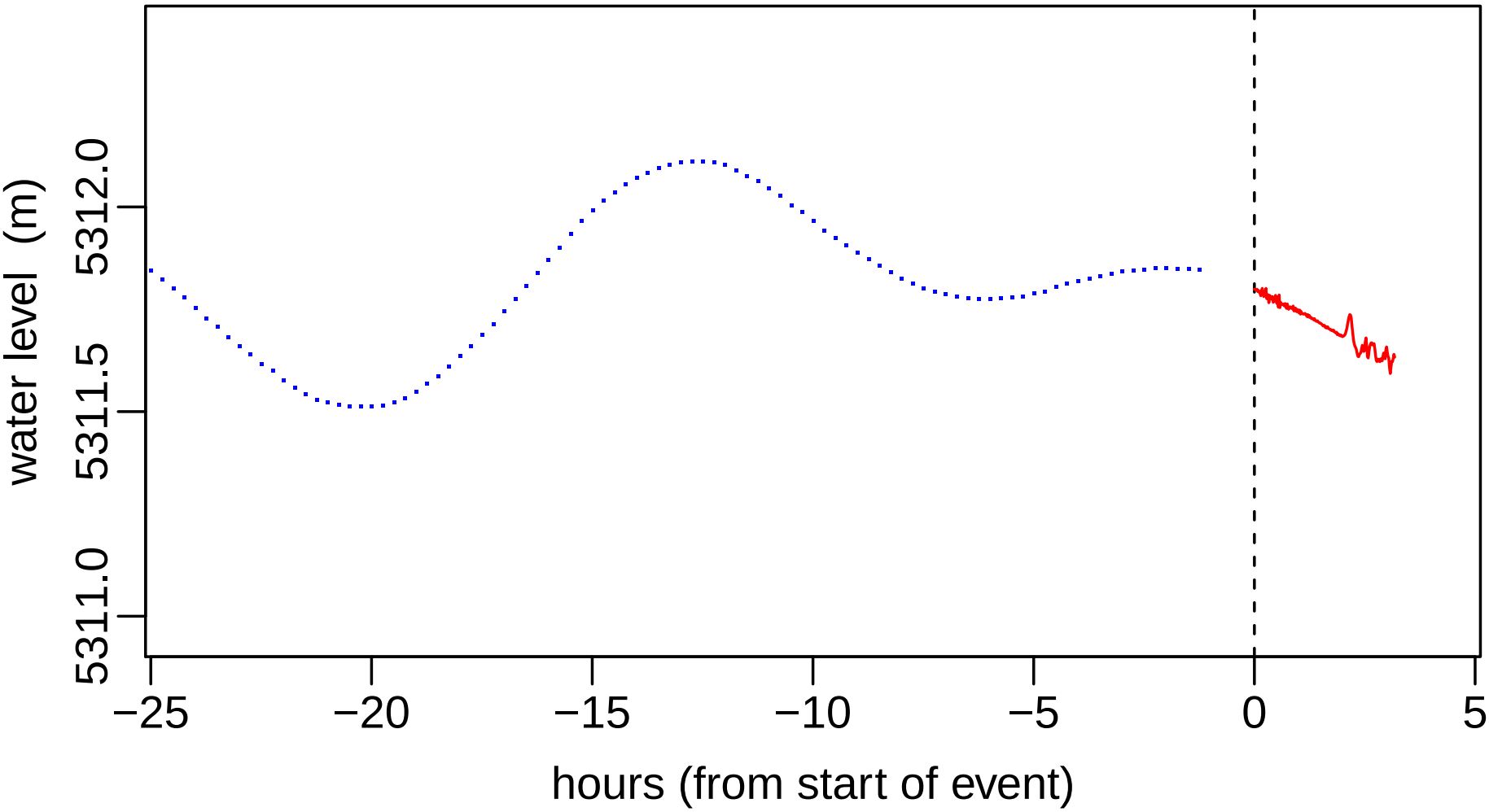
# Format of Data

- buoy records bottom pressure measurement every 15 sec (each is an average over 15 sec) – saved internally and retrieved when buoy serviced (typically 400 to 1000 days between servicings)
- 15-min stream: when nothing is going on, buoy transmits one 15-sec average every 15 min (most recent portion of stream displayed on NOAA Web site)
- 15-sec stream: when triggered by an event (typically seismic noise), buoy transmits $3\frac{3}{4}$ minute segment of 15-sec averages
- 1-min stream: during tsunami event, buoy transmits averages of four consecutive 15-sec averages (i.e., 1-min averages)
- due to transmission protocols, what is actually available during tsunami event can be complicated mixture of these streams (subject to change with deployment of 4th generation DART®'s)

# Buoy 21414 Data for Nov. 2006 Kuril Islands Event



# Buoy 21414 Data for Nov. 2006 Kuril Islands Event



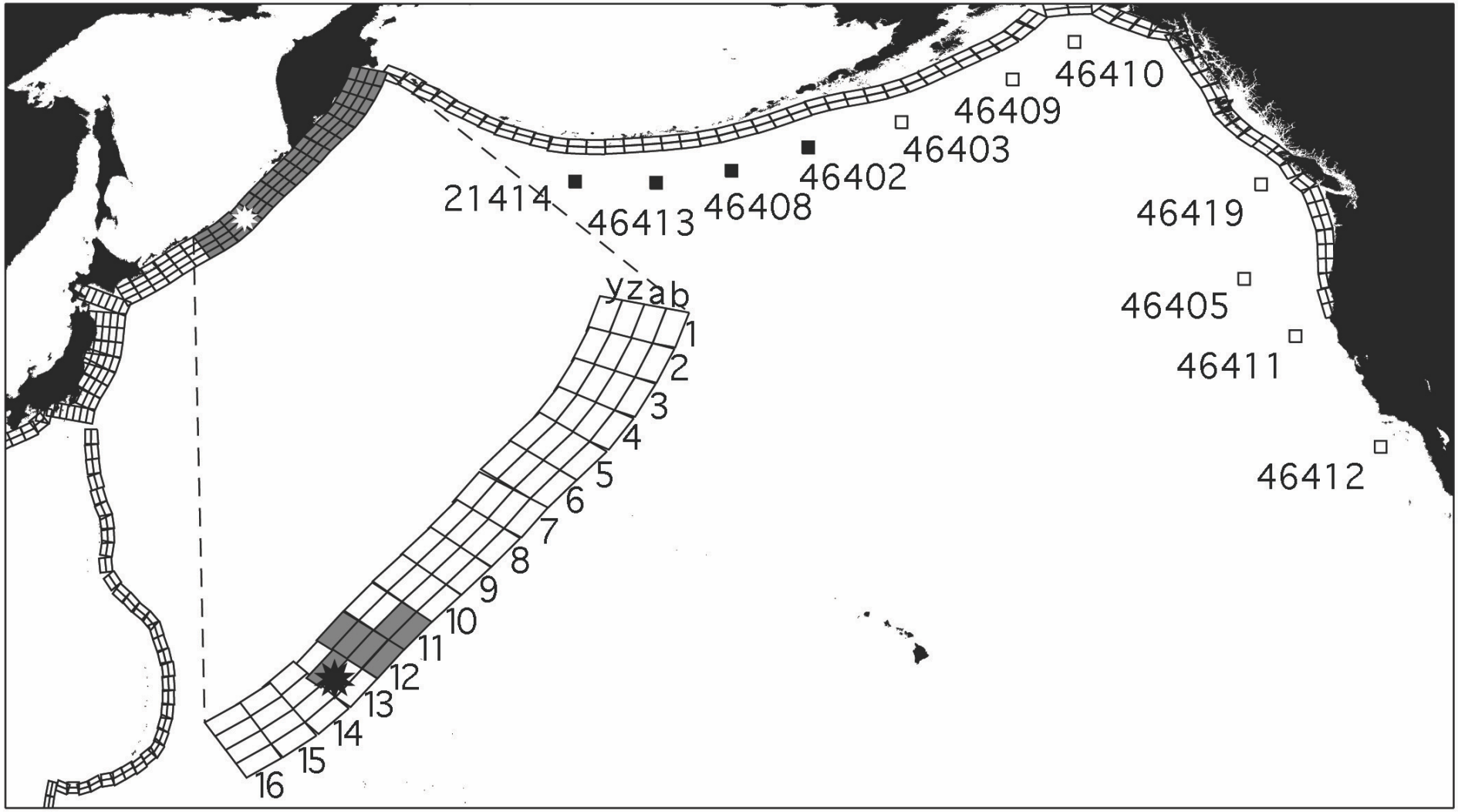
## Short-term Inundation Forecast for Tsunamis (SIFT)

- SIFT: computer application that uses DART® data and pre-computed geophysically-based predictors to assess magnitude of tsunami as event evolves – currently in use at Alaska & Hawaii warning centers, but being refined continually at NOAA Center for Tsunami Research
- statistical issues in processing data within SIFT include:
  1. predictor selection: currently done by hand by trained operators, but automation highly desired
  2. estimation of tsunami signal: fit predictors, keeping in mind confounding tidal component
  3. assessment of uncertainty in coastal inundation forecasts
- will focus on 1 and 2 in today's talk

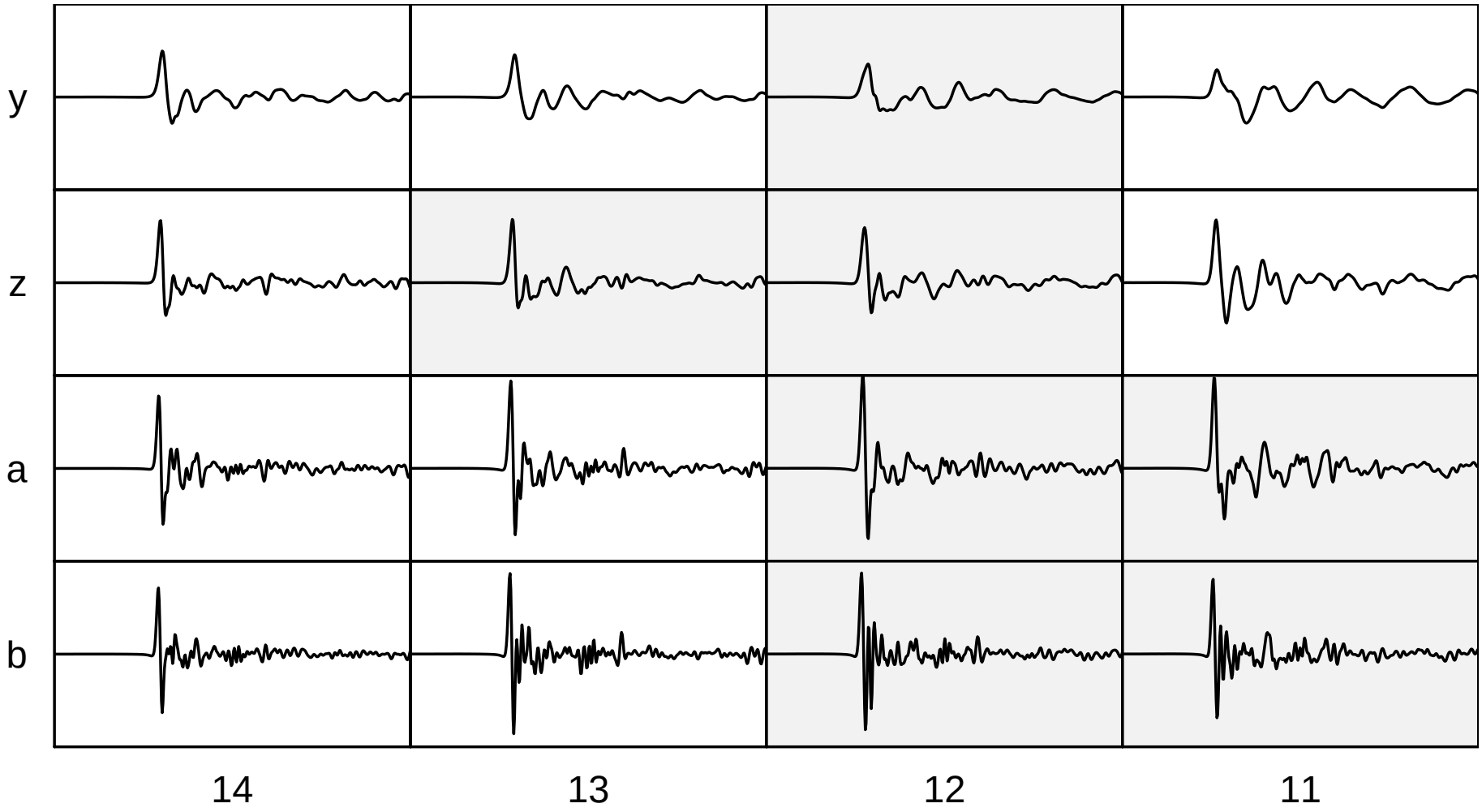
# Predictors for Tsunami Signals: I

- predictors for tsunami signals depend on both
  - location of seismic disturbance and
  - location of DART® buoy
- each predictor is associated with so-called 'unit source', which is a 100 km by 50 km section of ocean near a coastline
- predictors reside in precomputed database – one for each pairing of a particular unit source and particular buoy
- predictors describe what would be observed – *in absence of tidal effects* – at buoy if a moment magnitude $M_W = 7.5$ inverse-thrust fault earthquake were to originate from associated unit source (Okada, 1985)

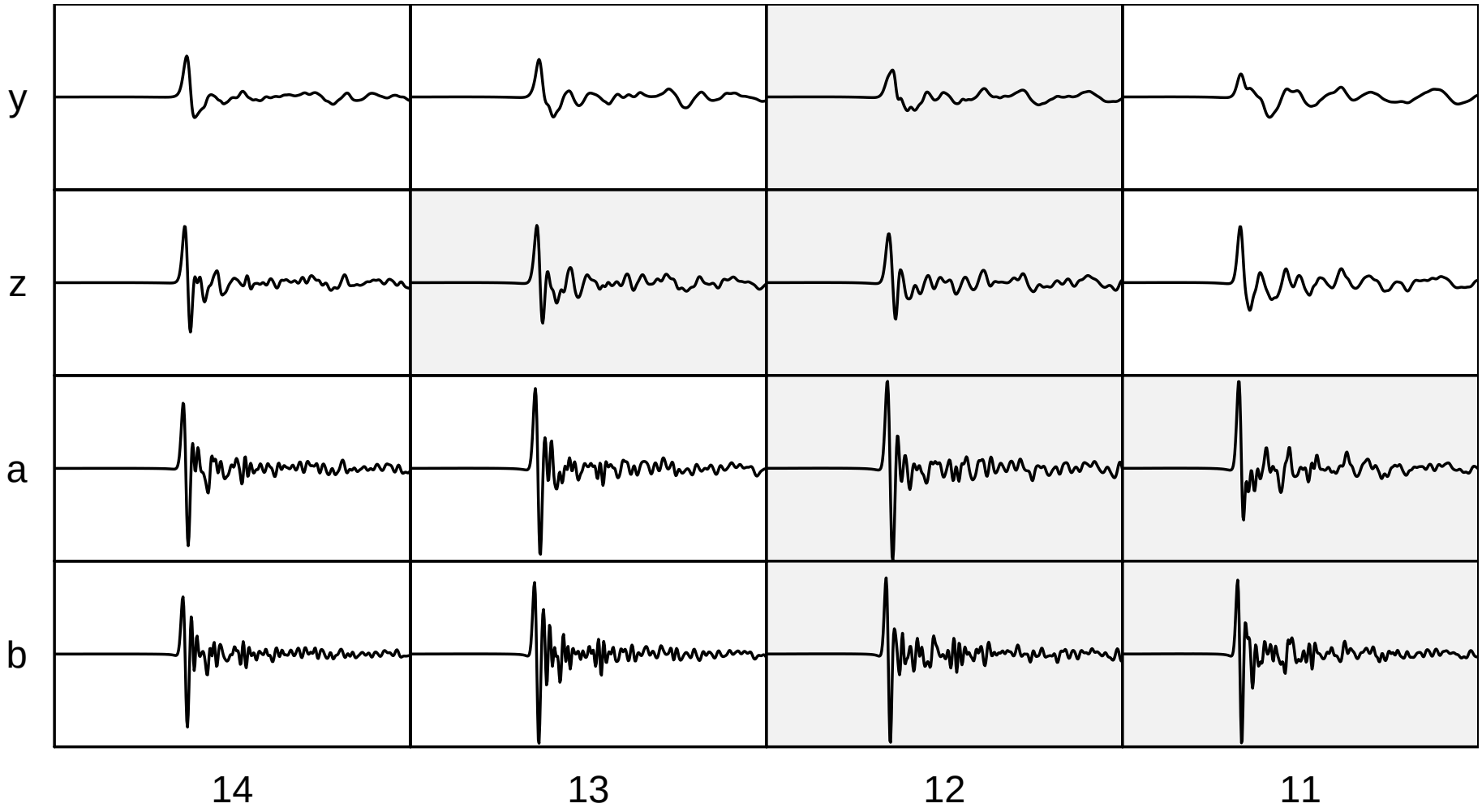
# Predictors for Nov. 2006 Kuril Islands Event



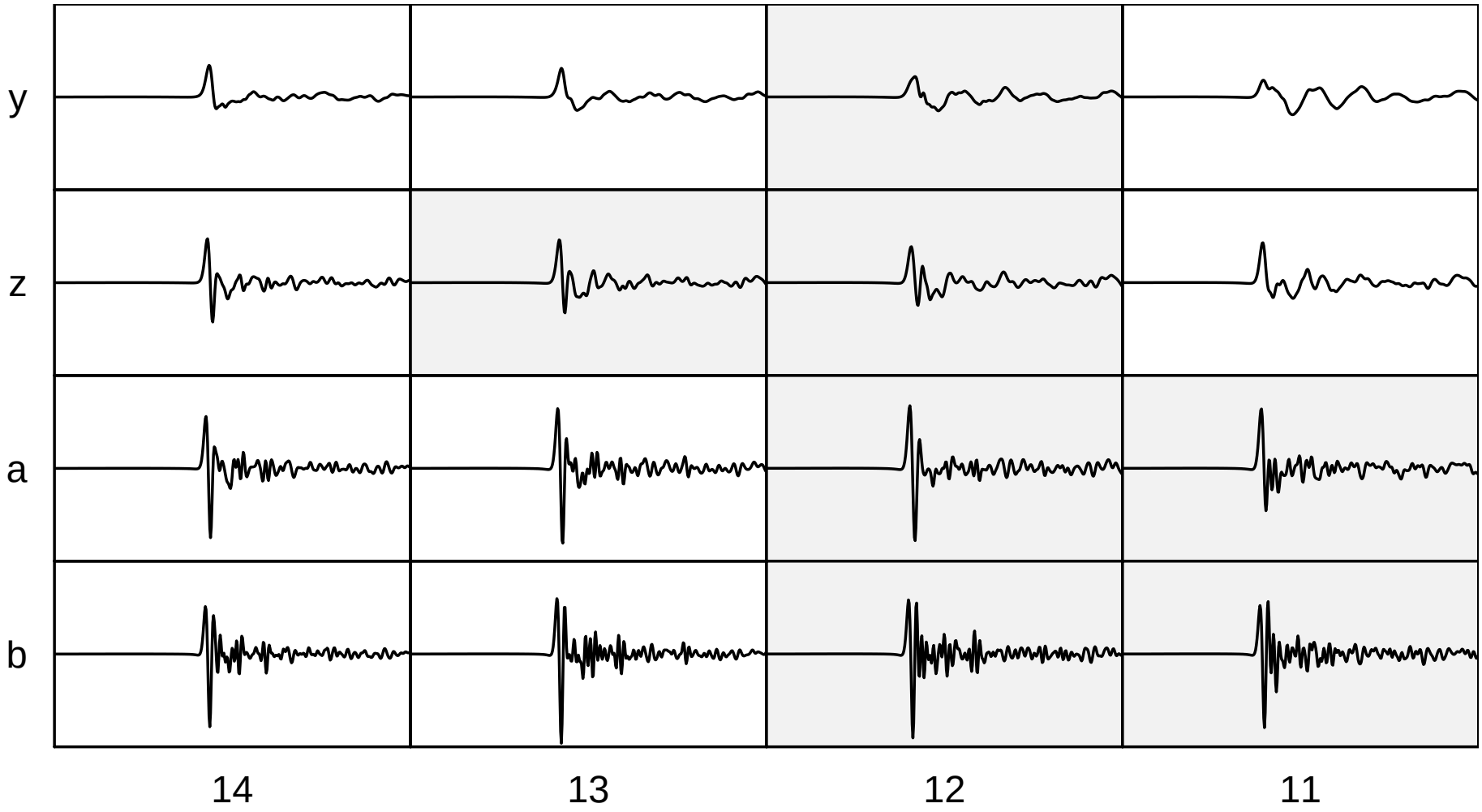
# Predictors for Buoy 21414 (0 to 8 hours)



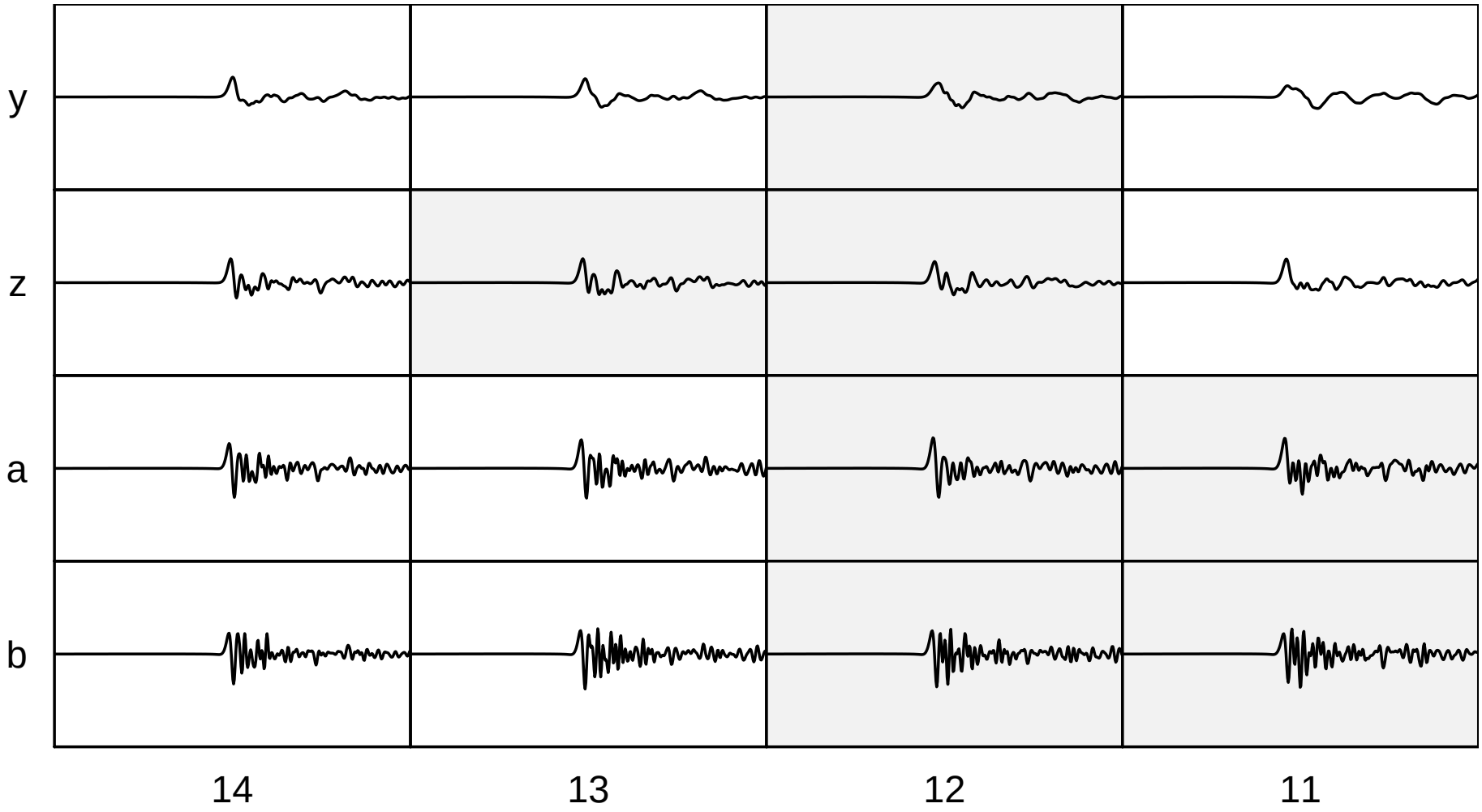
# Predictors for Buoy 46413 (0 to 8 hours)



# Predictors for Buoy 46408 (0 to 8 hours)



# Predictors for Buoy 46402 (0 to 8 hours)



# Predictors for Tsunami Signals: II

- for a given buoy, predictors from different unit sources resemble one another (will lead to collinearity concerns)
- let $\mathbf{g}_{j,k}$ be column vector containing predicted time series for $j$th buoy for $M_W = 7.5$ event occurring in $k$th unit source
- can account for event with $M_W$ differing from 7.5 by multiplying $\mathbf{g}_{j,k}$ by so-called source coefficient $\alpha_k \geq 0$
- prediction becomes $\alpha_k \mathbf{g}_{j,k}$

# Predictors for Tsunami Signals: III

- tsunami event often spread out over $K > 1$ unit sources
- physical considerations suggest that linearity is a reasonable assumption, but negative source coefficients don't make physical sense
- predictor for $j$th buoy thus becomes

$$\sum_{k=1}^{K} \alpha_k \mathbf{g}_{j,k} = G_j \boldsymbol{\alpha},$$

where

  - $K \geq 1$ is the number of unit sources under consideration
  - $G_j$ is a matrix whose columns are the $\mathbf{g}_{j,k}$ vectors
  - $\boldsymbol{\alpha}$ is a column vector with source coefficients $\alpha_k \geq 0$

# Models for DART® Buoy Data

- let $\mathbf{d}_j$ be column vector with time series collected by $j$th buoy *but after effect of tides has been removed* (more on this later!)
- model for detided data is

$$\mathbf{d}_j = G_j\boldsymbol{\alpha} + \mathbf{e}_j,$$

  where $\mathbf{e}_j$ is a stochastic vector of errors (more on these later)
- if we have relevant data from $J$ buoys in all, can create multi-buoy model by stacking components of individual models:

$$\mathbf{d} = G\boldsymbol{\alpha}+\mathbf{e}, \text{ where } \mathbf{d} = \begin{bmatrix} \mathbf{d}_1 \\ \vdots \\ \mathbf{d}_J \end{bmatrix}, \; G = \begin{bmatrix} G_1 \\ \vdots \\ G_J \end{bmatrix} \text{ and } \mathbf{e} = \begin{bmatrix} \mathbf{e}_1 \\ \vdots \\ \mathbf{e}_J \end{bmatrix}$$

# Error Component in Model: I

- autocorrelation is typical in time series – cannot realistically assume that errors terms in $\mathbf{e}_j$ are uncorrelated
- will assume $\mathbf{e}_j$ is multivariate Gaussian with known positive-definite covariance matrix $\sigma^2\Sigma_{\mathbf{e}_j}$ – thus

$$\mathbf{e}_j \sim \mathcal{N}(\mathbf{0}, \sigma^2\Sigma_{\mathbf{e}_j}),$$

  where $\sigma^2 > 0$ is a scaling factor
- Cholesky decomposition allows construction of matrix $\Sigma_{\mathbf{e}_j}^{-1/2}$ such that

$$\Sigma_{\mathbf{e}_j}^{-1/2}\Sigma_{\mathbf{e}_j}^{-1/2} = \Sigma_{\mathbf{e}_j}^{-1},$$

  where $\Sigma_{\mathbf{e}_j}^{-1}$ is the inverse of $\Sigma_{\mathbf{e}_j}$

# Error Component in Model: II

- leads to transformed model

$$\Sigma_{\mathbf{e}_j}^{-1/2}\mathbf{d}_j = \Sigma_{\mathbf{e}_j}^{-1/2}G_j\boldsymbol{\alpha} + \Sigma_{\mathbf{e}_j}^{-1/2}\mathbf{e}_j$$

  with independent & identically distributed (IID) errors $\Sigma_{\mathbf{e}_j}^{-1/2}\mathbf{e}_j$

- ordinary least squares (LS) estimator of $\boldsymbol{\alpha}$ in transformed model is generalized LS estimator for original model $\mathbf{d}_j = G_j\boldsymbol{\alpha} + \mathbf{e}_j$

- two additional simplifying assumptions
  - matrix $\Sigma_{\mathbf{e}_j}$ dictated by first-order autoregressive process with unit lag autocorrelation $\phi$ (can estimate via linear regression, and $\Sigma_{\mathbf{e}_j}^{-1/2}$ is nonzero only in diagonal & 1st subdiagonal)
  - errors in one buoy are independent of those in any other buoy

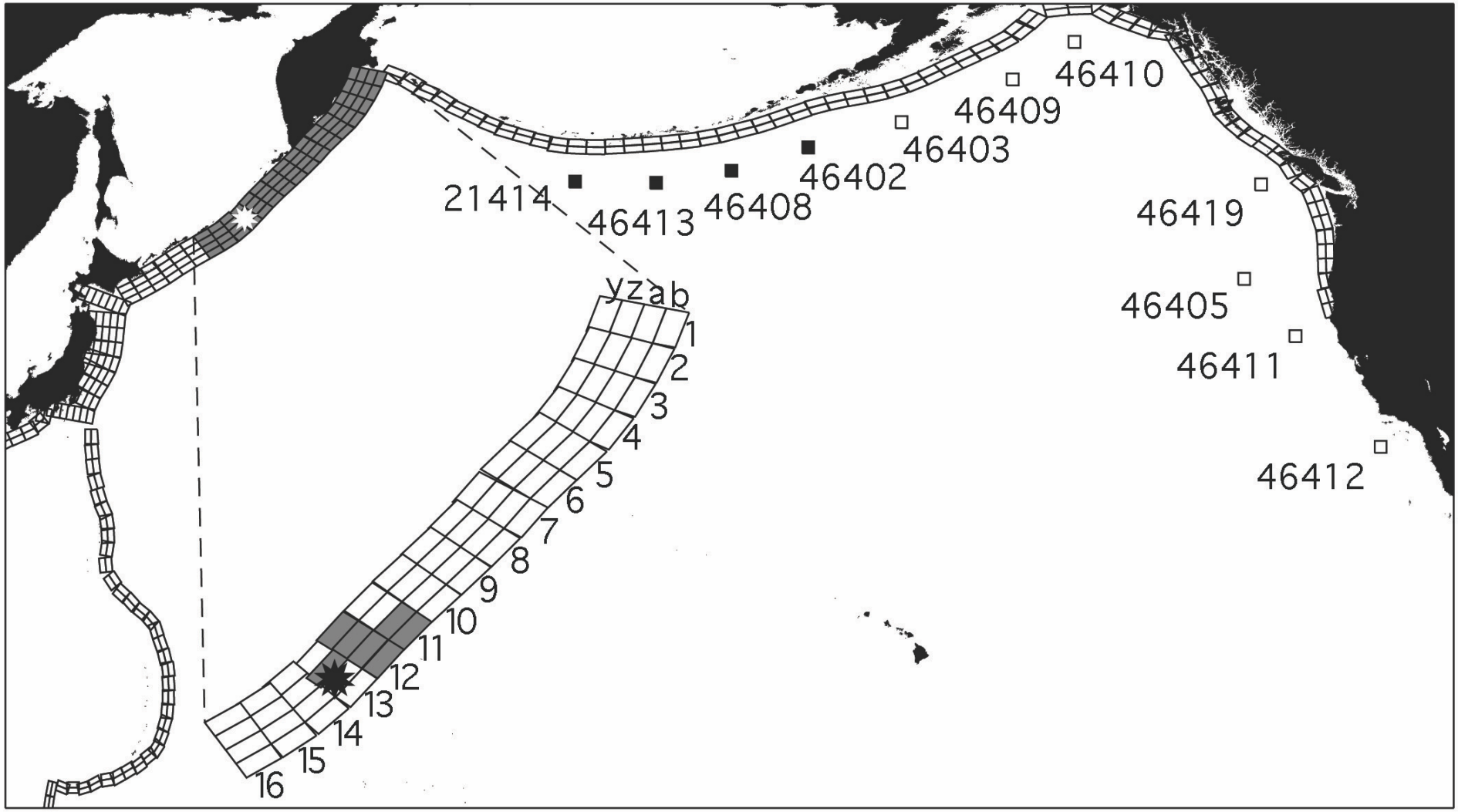
- leads to $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma^2\Sigma_{\mathbf{e}})$ in multi-buoy model $\mathbf{d} = G\boldsymbol{\alpha} + \mathbf{e}$, where $\Sigma_{\mathbf{e}}$ is block diagonal (decorrelation matrix $\Sigma_{\mathbf{e}}^{-1/2}$ is also)

# Selection of Predictors and $\alpha$ Estimation: I

- given
  - transformed data $\tilde{\mathbf{d}} = \Sigma_{\mathbf{e}}^{-1/2}\mathbf{d}$
  - transformed predictors $\widetilde{G} = \Sigma_{\mathbf{e}}^{-1/2}G$
  - model $\tilde{\mathbf{d}} = \widetilde{G}\boldsymbol{\alpha} + \tilde{\mathbf{e}}$ with IID Gaussian errors

  task is to estimate source coefficients $\boldsymbol{\alpha}$
- *lots* of potential predictors – one for each unit source in relevant subduction zone(s)
- $K = 22 \times 4 = 88$ unit sources in Kamchatka–Kuril–Japan subduction zone where earthquake generating Kuril Islands event occurred

# Predictors for Nov. 2006 Kuril Islands Event



# Selection of Predictors and $\alpha$ Estimation: II

- want solution $\hat{\boldsymbol{\alpha}}$ to $\tilde{\mathbf{d}} = \widetilde{G}\boldsymbol{\alpha} + \tilde{\mathbf{e}}$ to be
  - *sparse*: many entries of $\hat{\boldsymbol{\alpha}}$ should be equal to zero (unlikely for earthquake to be spread out over an entire subduction zone)
  - *structured*: earthquakes are typically spatially localized, so nonzero entries of $\hat{\boldsymbol{\alpha}}$ should correspond to unit sources close to one another
  - *nonnegative*: $\hat{\alpha}_k < 0$ does not make physical sense

# Selection of Predictors and $\alpha$ Estimation: III

- ordinary LS (OLS) estimator with nonnegativity constraints, i.e., estimator minimizing

$$\|\tilde{\mathbf{e}}\|_2^2 \stackrel{\text{def}}{=} \sum_{j=1}^{J}\sum_{n=1}^{N_j} e_{j,n}^2 = \|\tilde{\mathbf{d}} - \widetilde{G}\boldsymbol{\alpha}\|_2^2 \text{ subject to } \boldsymbol{\alpha} \geq \mathbf{0},$$

  imposes some degree of sparsity due to nonnegativity constraint, but can easily end up overfitting data if $K$ is large

  – no reason for constrained OLS estimator to be structured

# Selection of Predictors and $\alpha$ Estimation: IV

- to achieve sparsity, consider *lasso* proposed by Tibshirani (1996)
- constrained version of lasso yields estimator minimizing

$$\|\tilde{\mathbf{d}} - \widetilde{G}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1 \text{ subject to } \boldsymbol{\alpha} \geq \mathbf{0},$$

  where

$$\|\boldsymbol{\alpha}\|_1 \stackrel{\text{def}}{=} \sum_{k=1}^{K} |\alpha_k| = \sum_{k=1}^{K} \alpha_k$$

  and $\lambda\|\boldsymbol{\alpha}\|_1$ is the so-called $\ell_1$ penalty
- if $\lambda = 0$, lasso solution for $\boldsymbol{\alpha}$ reduces to constrained OLS
- if $\lambda = \infty$, lasso solution is $\hat{\boldsymbol{\alpha}} = \mathbf{0}$
- as $\lambda$ decreases from $\infty$, solution $\hat{\boldsymbol{\alpha}}$ becomes less sparse

# Selection of Predictors and $\alpha$ Estimation: V

- note: can use `R` package **`penalized`** to find lasso solution
- lasso solution $\hat{\boldsymbol{\alpha}}$ is a *biased* estimator of $\boldsymbol{\alpha}$, with bias being towards zero . . . yikes!!!
- $\boldsymbol{\alpha}$ estimates are used with geophysical models to forecast wave heights in open ocean near coastal community of interest
- in turn, open-ocean forecasts are used as initial conditions for run-up models that yield forecasts of coastal inundation
- downward bias in $\hat{\boldsymbol{\alpha}} \Longrightarrow$ downward bias in wave height forecasts $\Longrightarrow$ downward bias in coastal inundation forecasts

  – cannot sell this idea to those in charge of issuing warnings!
- will thus use lasso to identify which unit sources to use, and then estimate $\boldsymbol{\alpha}$ via constrained OLS

# Selection of Predictors and $\alpha$ Estimation: VI

- while lasso gives sparse solutions, no guarantee these solutions will be structured (same problem as with constrained OLS)
- proposal: restrict candidate unit sources to a localized region
- in example to follow, will consider $4 \times 3$ blocks of unit sources
- two-step procedure: for a given $4 \times 3$ localized region

  (1) use lasso with unit sources restricted to localized region

  (2) choose $\lambda$ tuning parameter (as described below) to automatically select unit sources within localized region
- repeat two-step procedure for all possible $4 \times 3$ regions
- strategy yields a localized solution: unit sources in the final model are contained within a single $4 \times 3$ region.

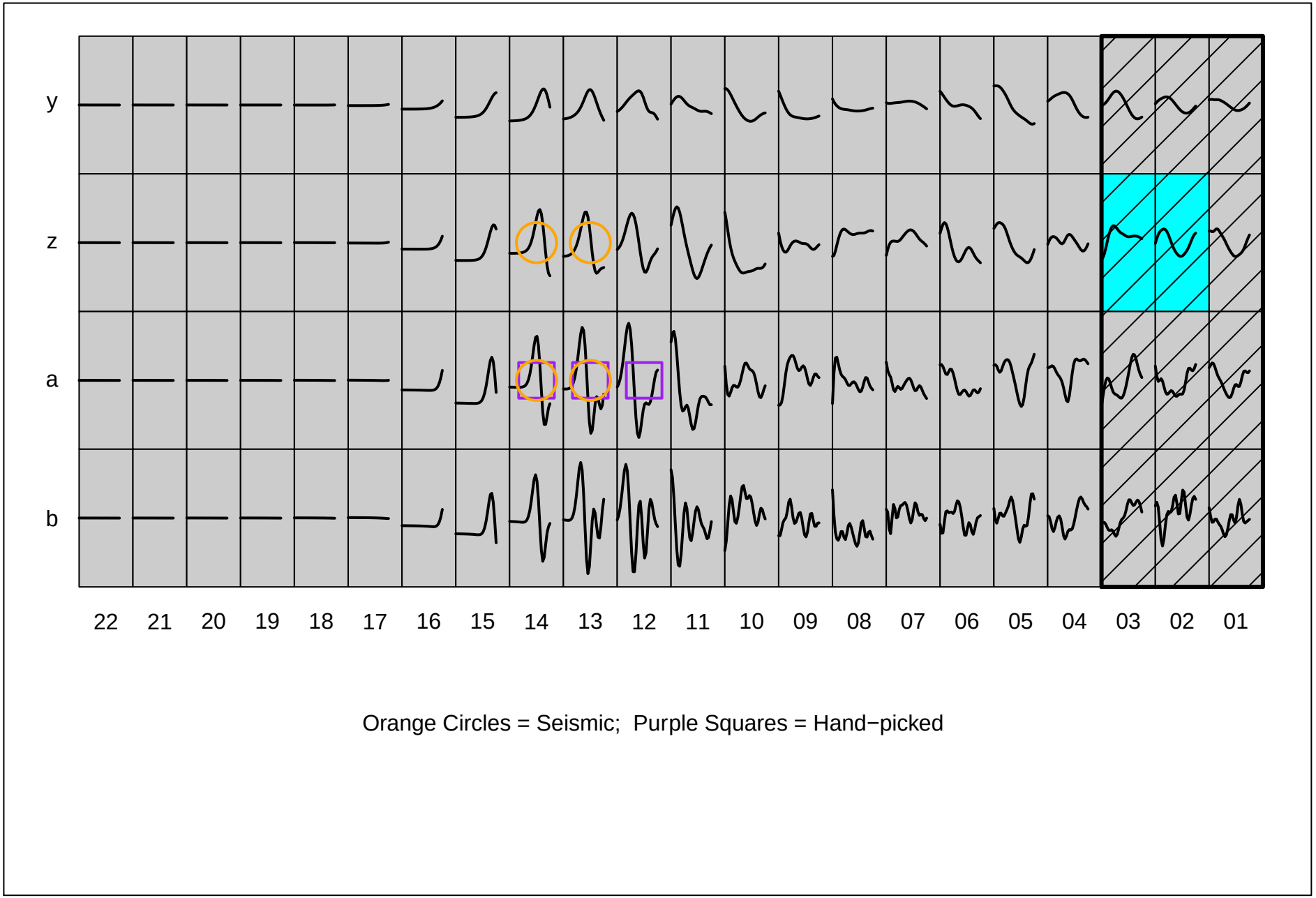
# Selection of Predictors and $\alpha$ Estimation: VII

- use Akaike information criterion (AIC) both to choose $\lambda$ within each localized region and to choose between regions
- intuition: could measure model's performance by residual sums of squares $\text{RSS} \stackrel{\text{def}}{=} \|\tilde{\mathbf{d}} - \widetilde{G}\hat{\boldsymbol{\alpha}}\|_2^2$, but this strategy calls for using data both to fit model and to measure its performance
- dual use of data leads to over-fitting ('curve matching'): chooses too low a $\lambda$, which translates into too many unit sources
- AIC fixes problem by adding a complexity penalty, a function of $N$ (# of data points) and $K'$ (# of selected unit sources):

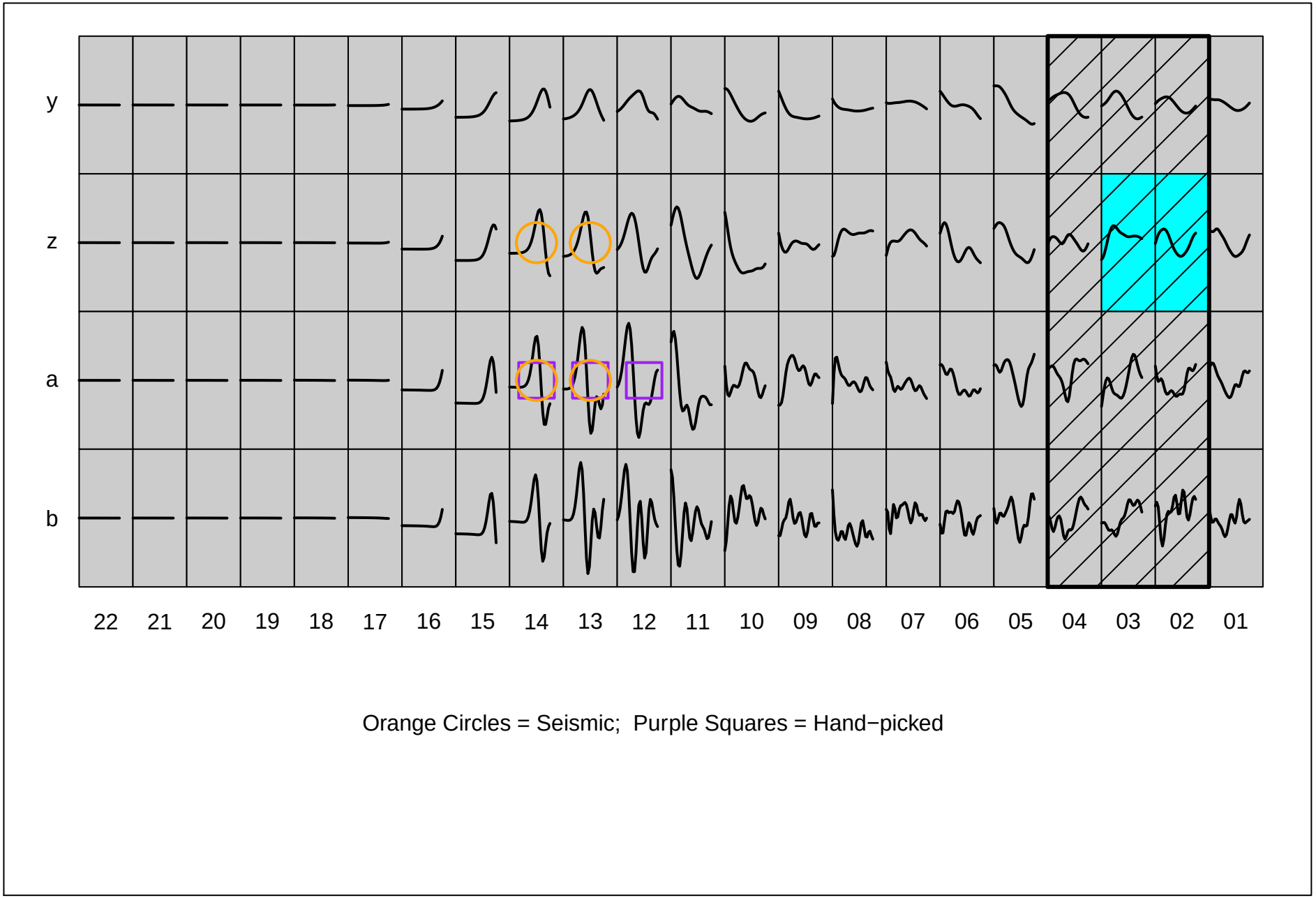
$$\text{AIC} \stackrel{\text{def}}{=} \log(\text{RSS}) + \text{Complexity Penalty}(N, K')$$

- strategy is to choose $\lambda$ yielding lowest AIC
- demo: sweeping lasso for Kuril Islands event with buoy 21414

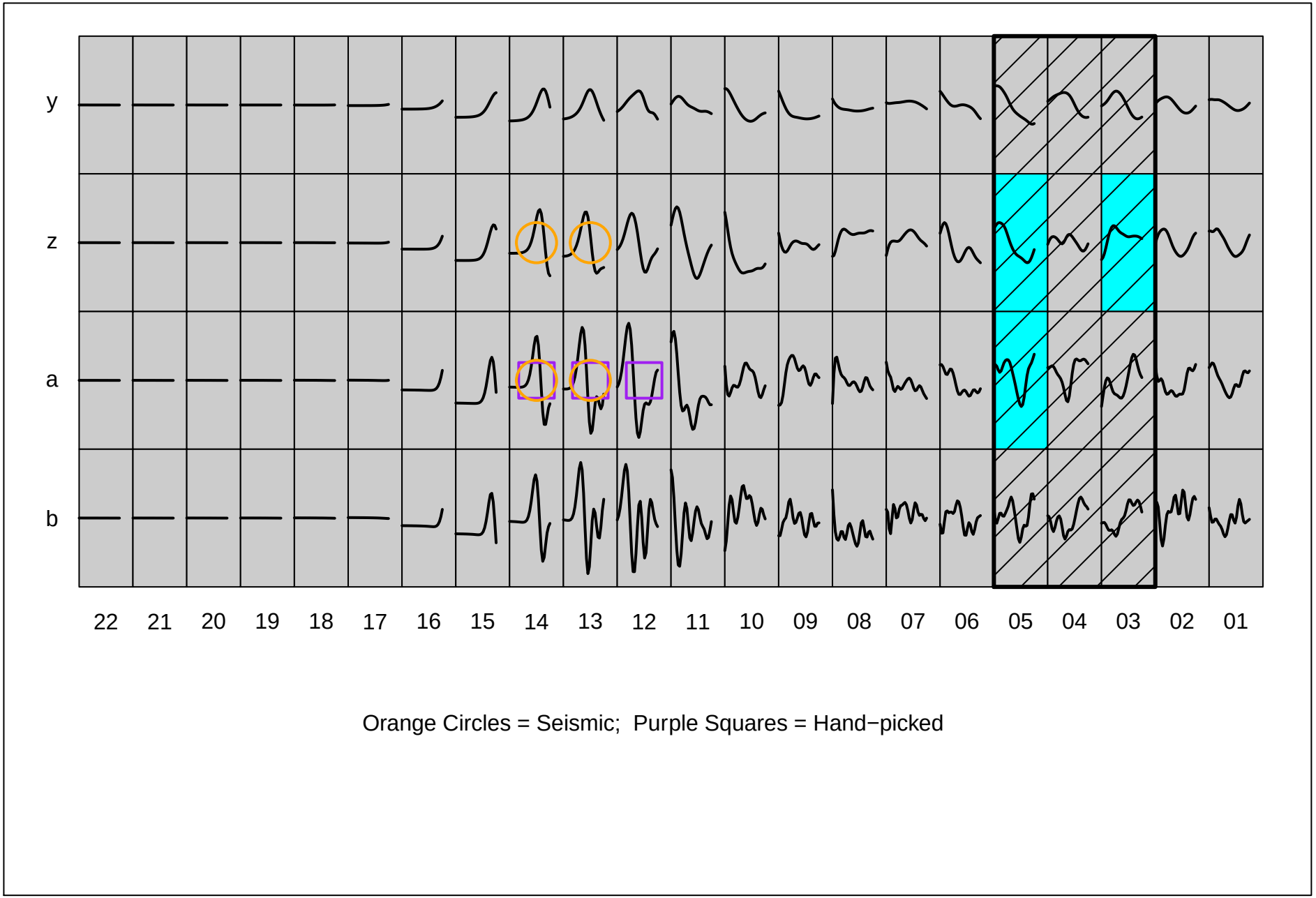
# Buoy 21414: AIC Score −128.4219



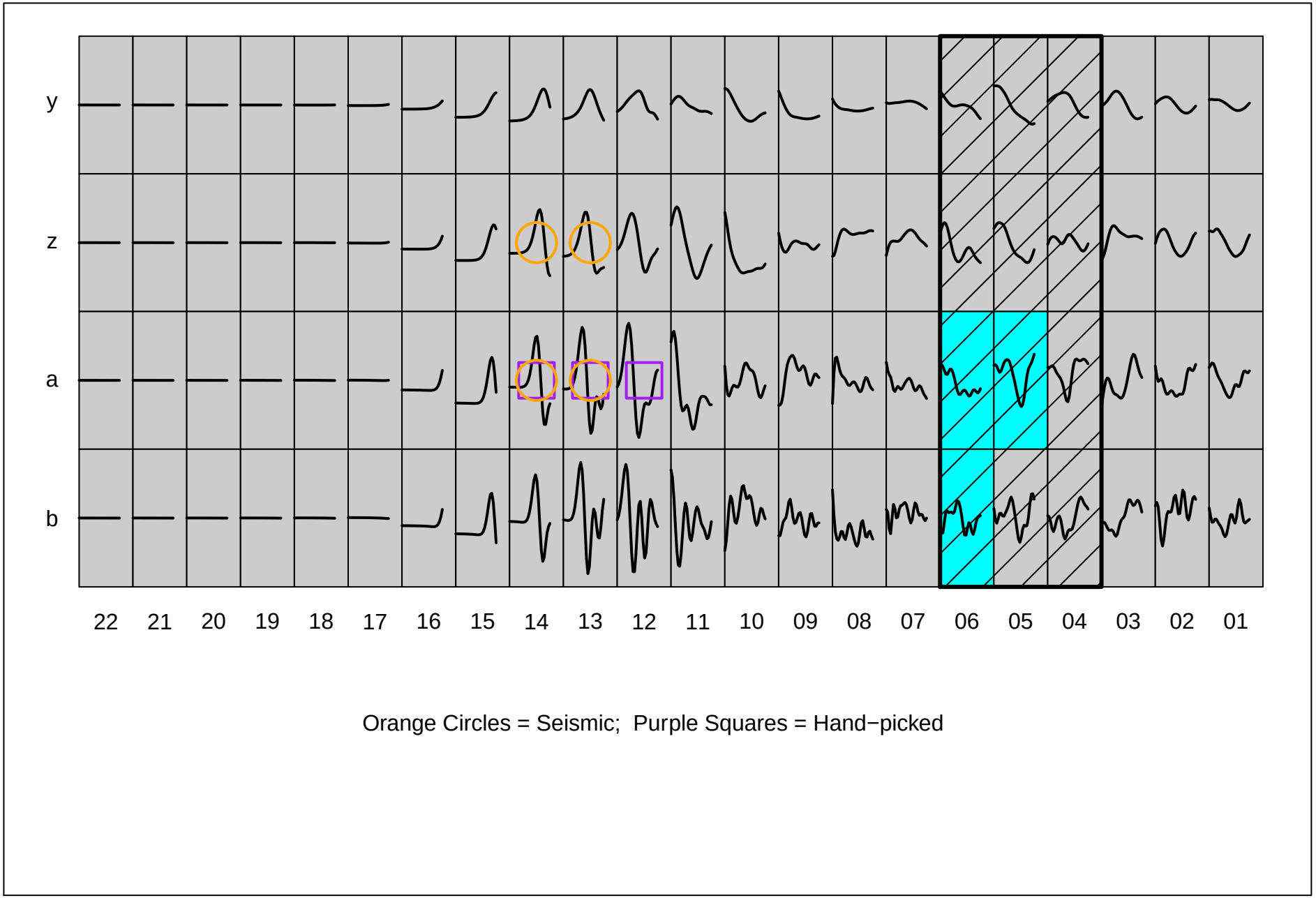
# Buoy 21414: AIC Score −128.4219



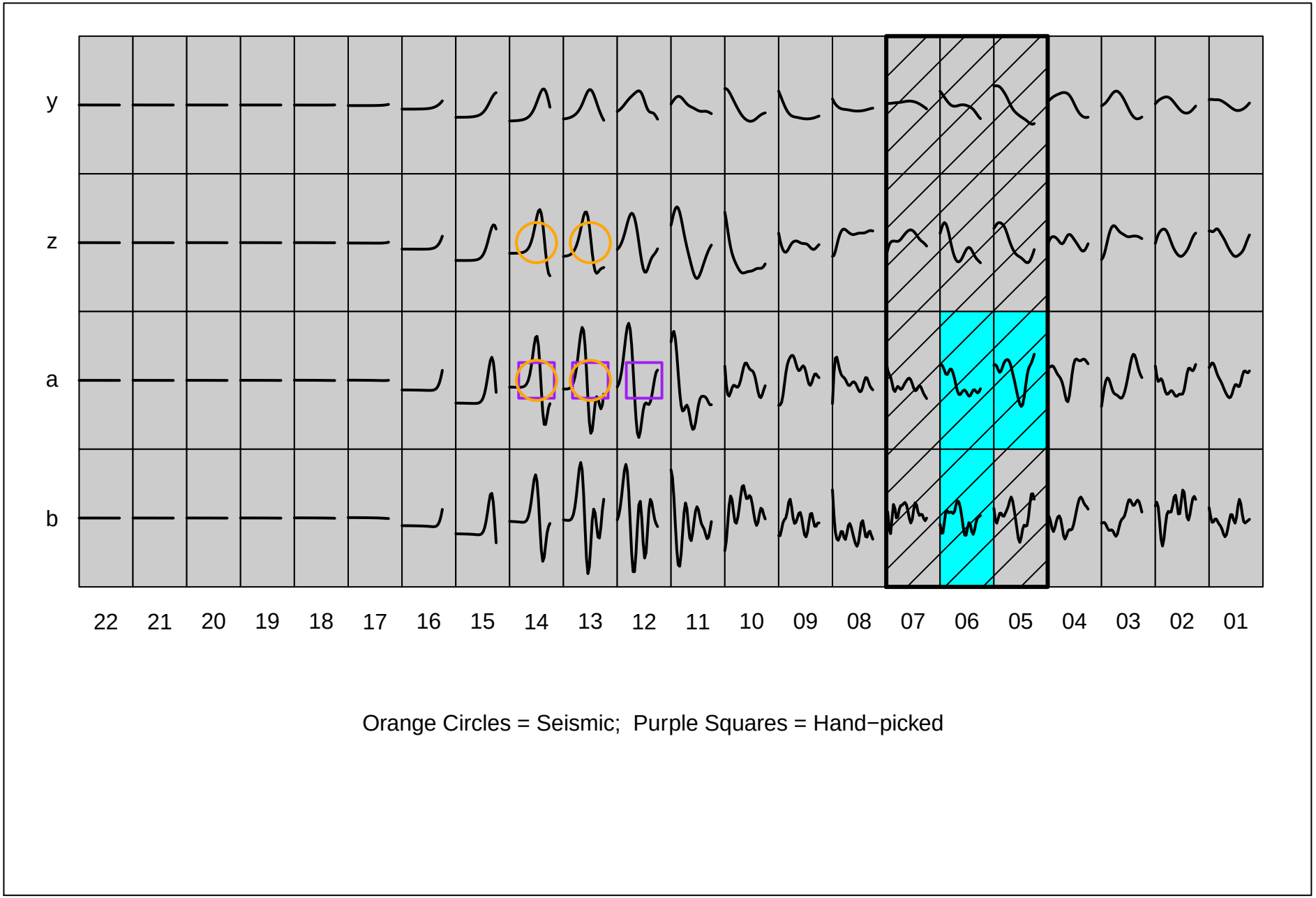
# Buoy 21414: AIC Score −131.5706



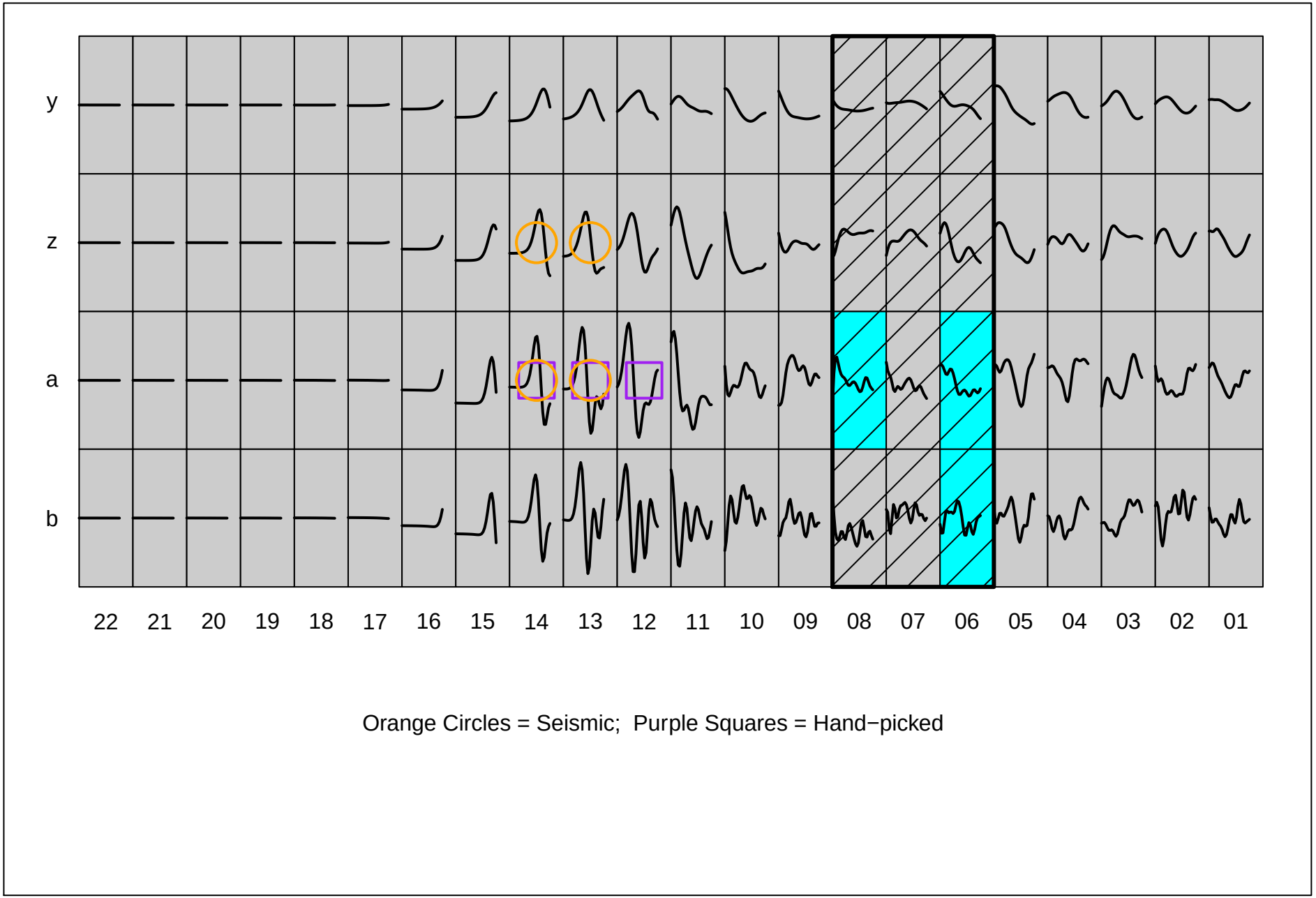
# Buoy 21414: AIC Score −134.0082



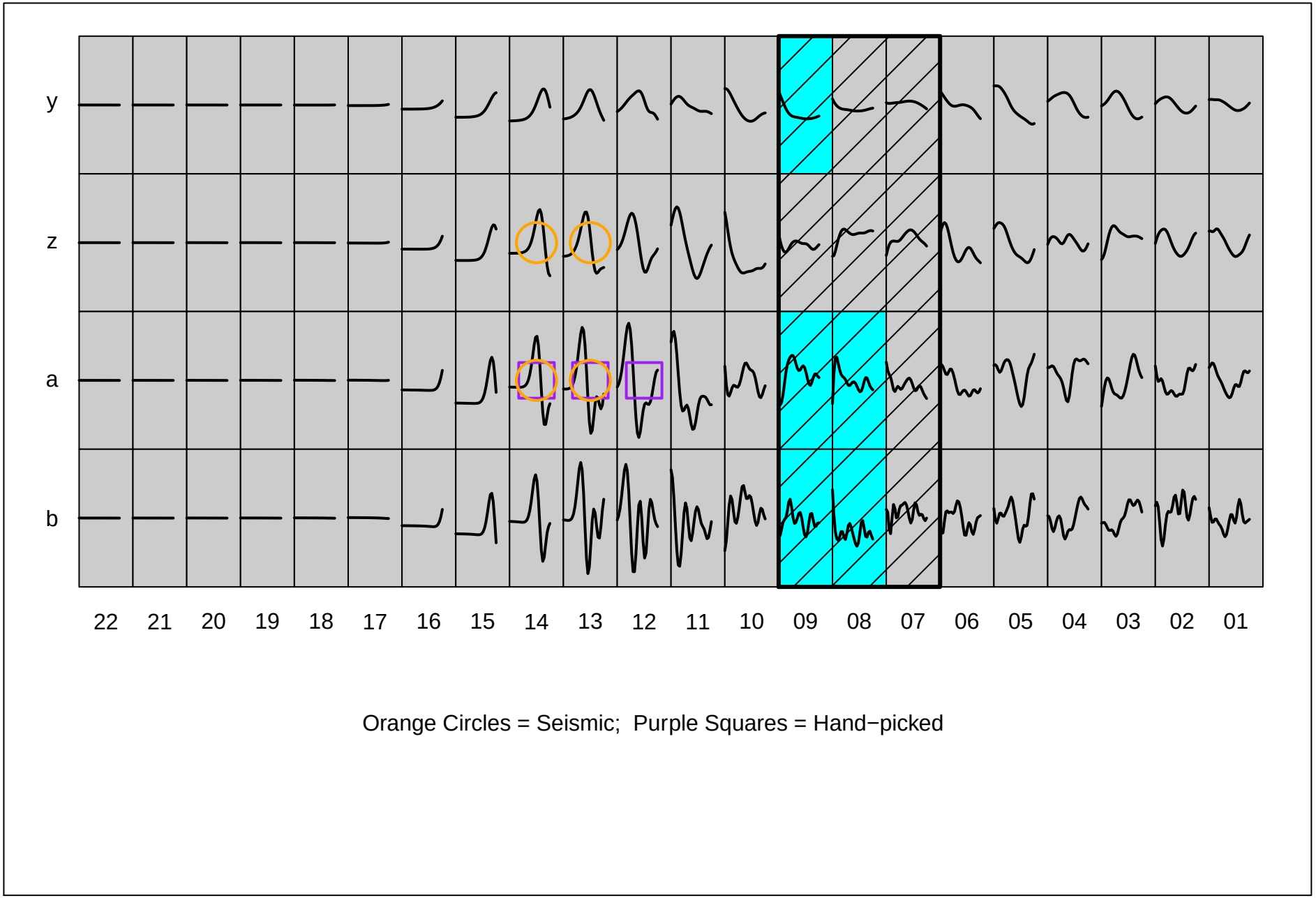
# Buoy 21414: AIC Score −134.0082



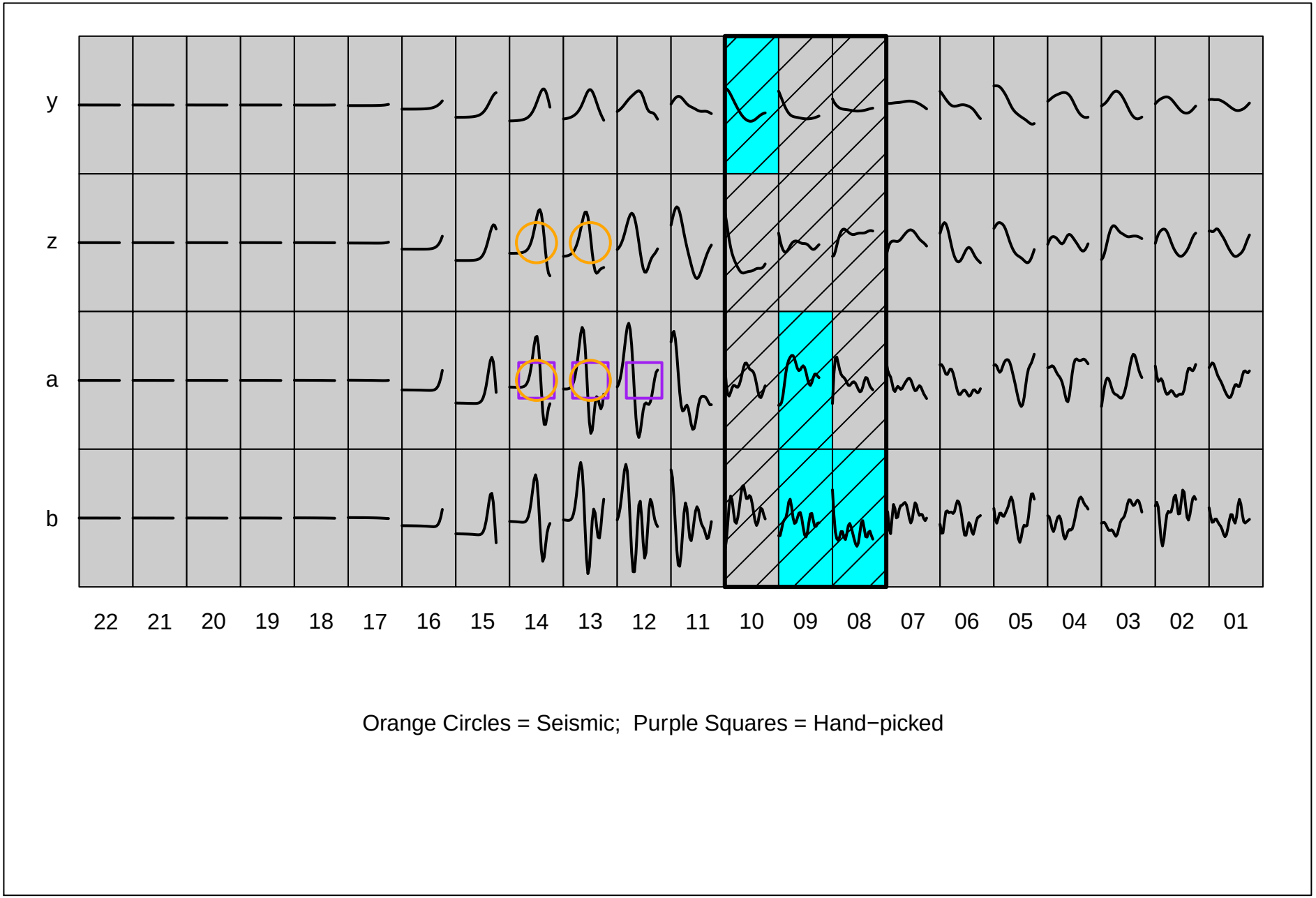
# Buoy 21414: AIC Score −136.5252



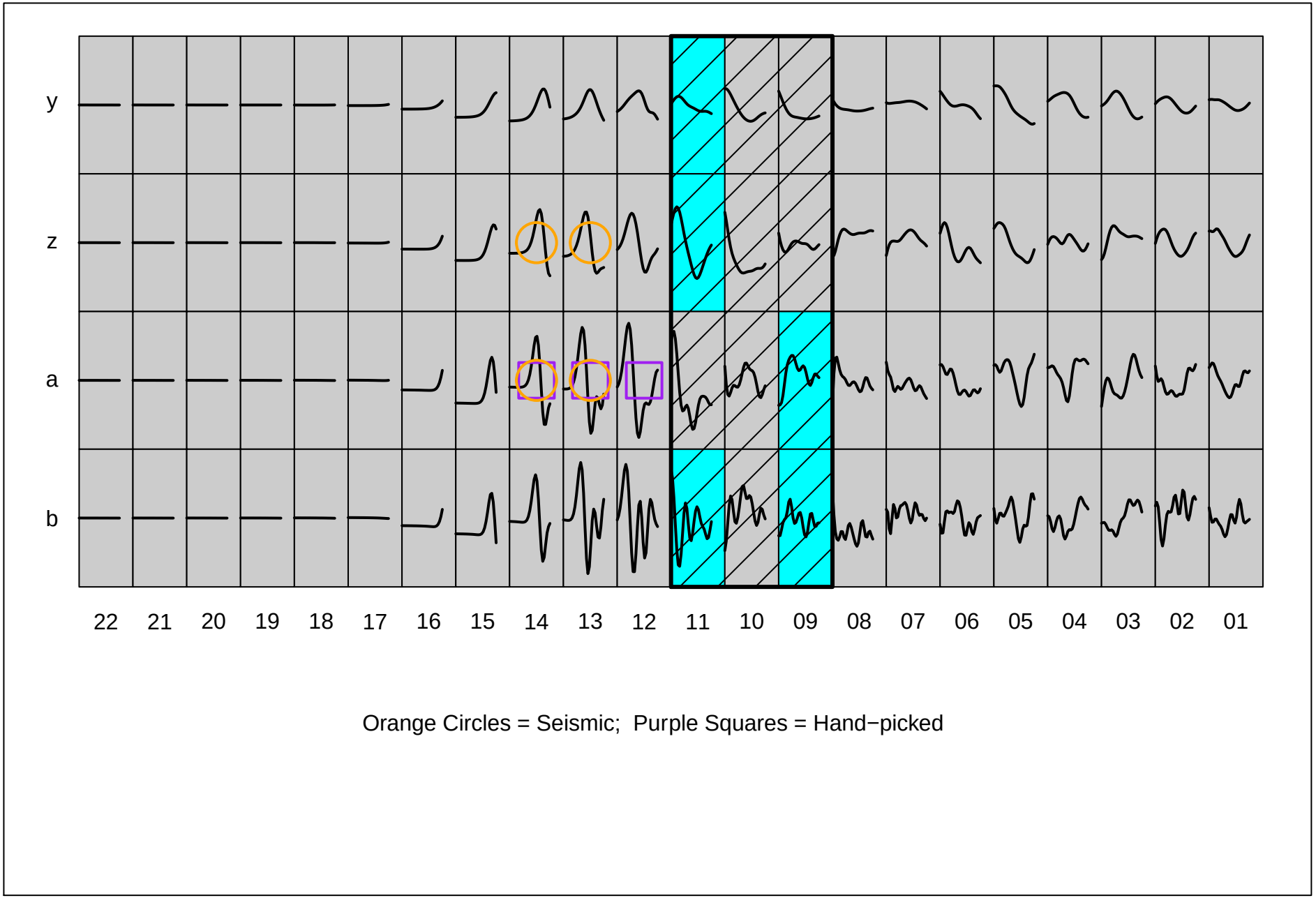
# Buoy 21414: AIC Score −139.4103



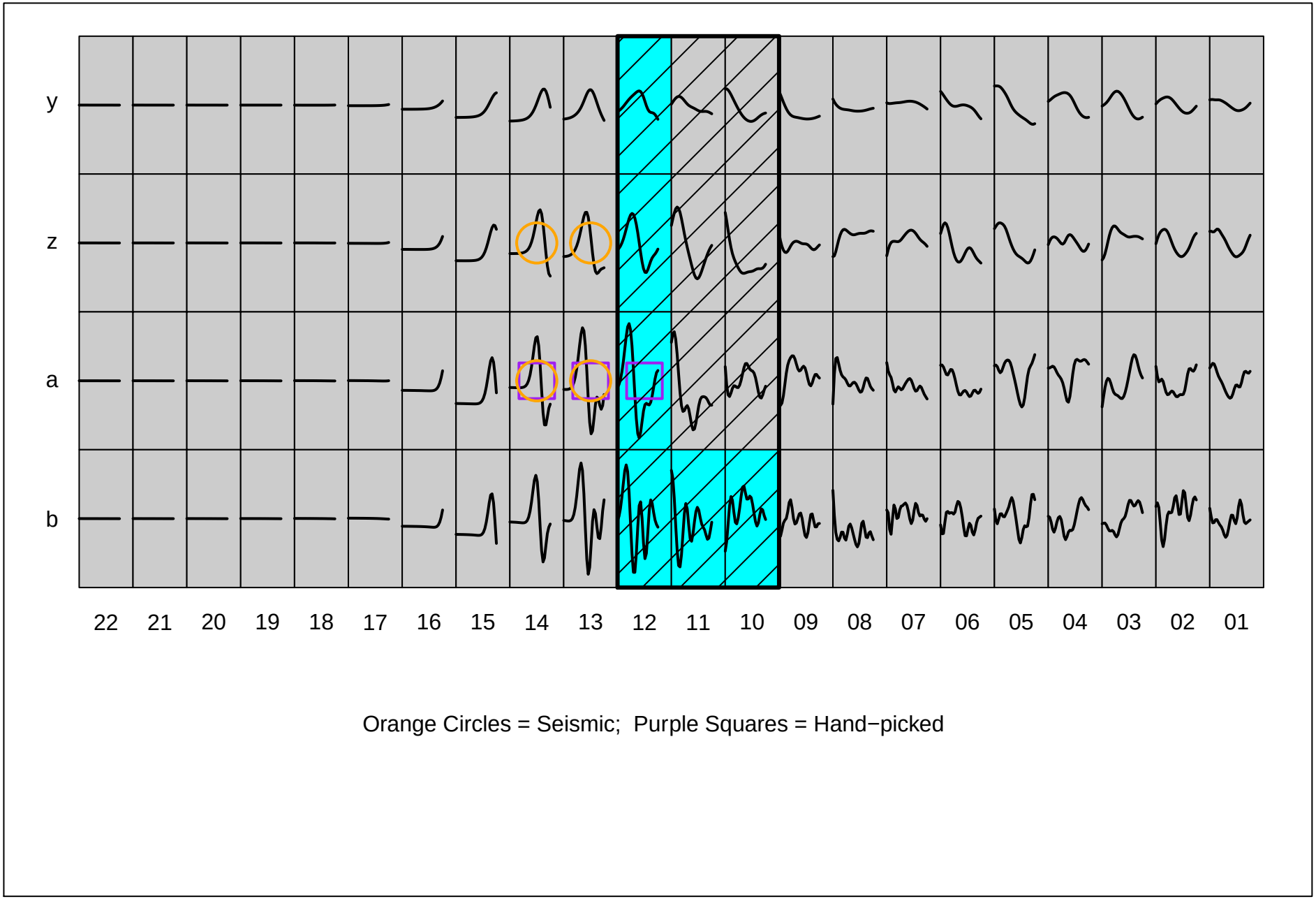
# Buoy 21414: AIC Score −174.4669



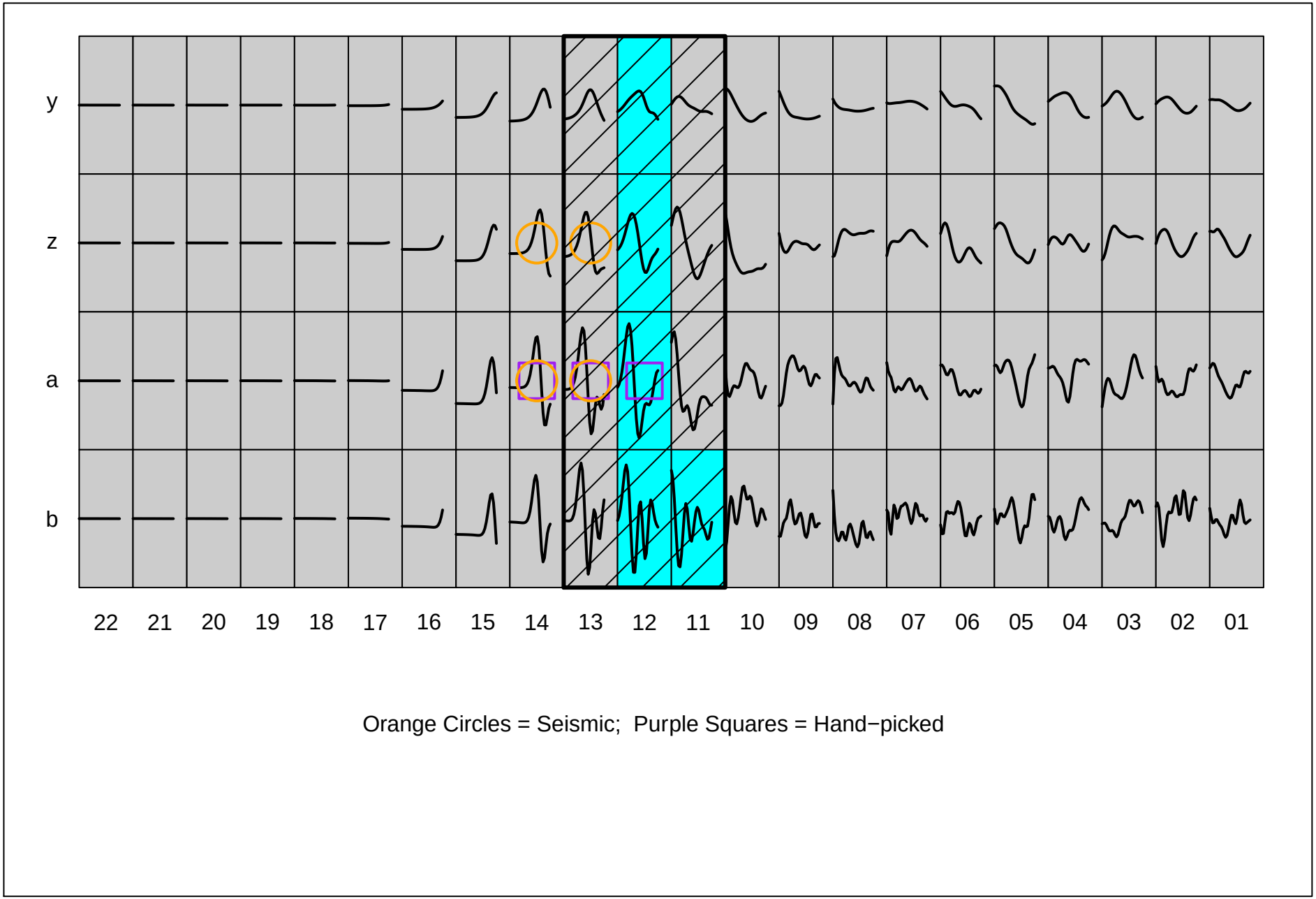
# Buoy 21414: AIC Score −202.4663



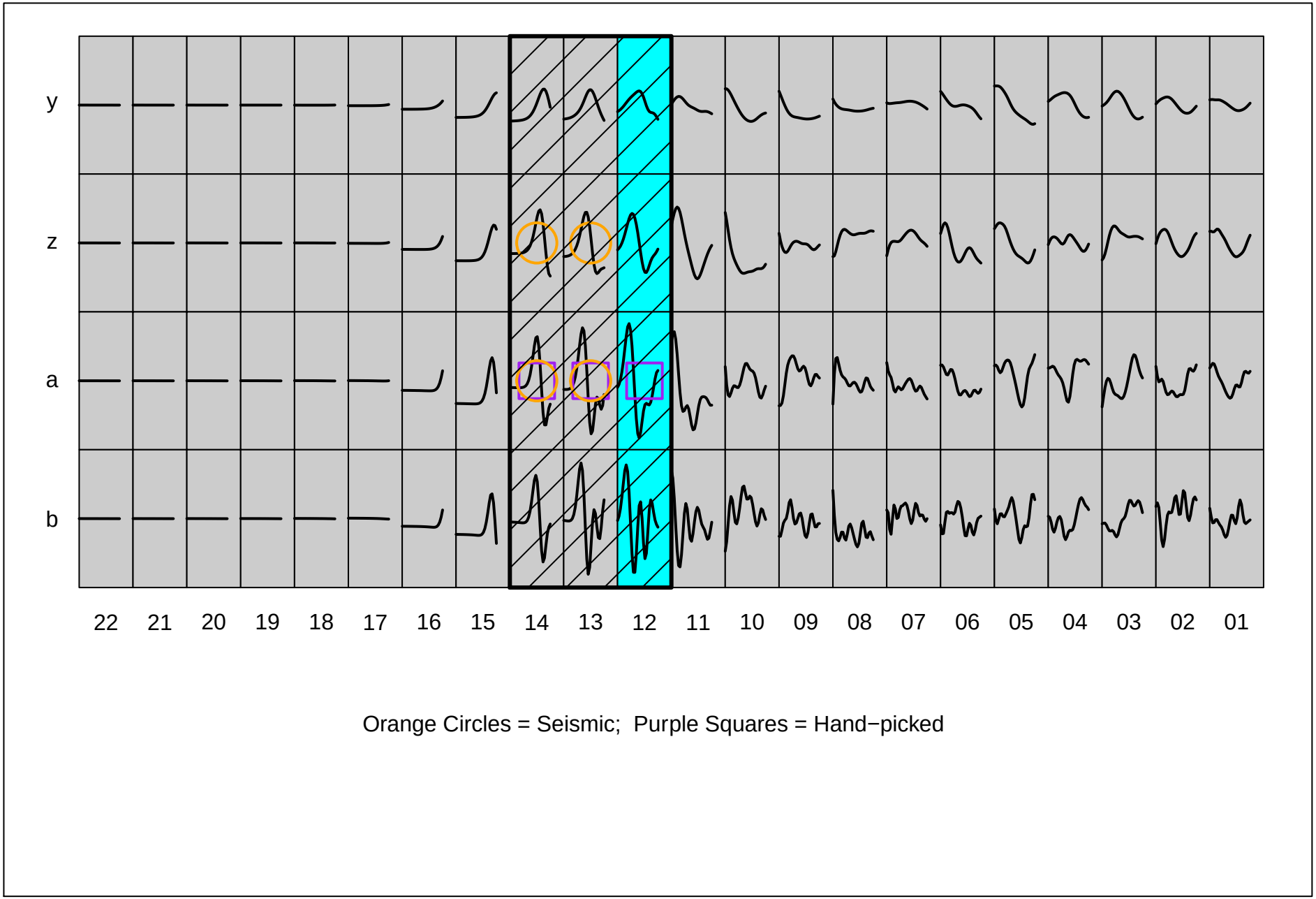
# Buoy 21414: AIC Score −228.3352



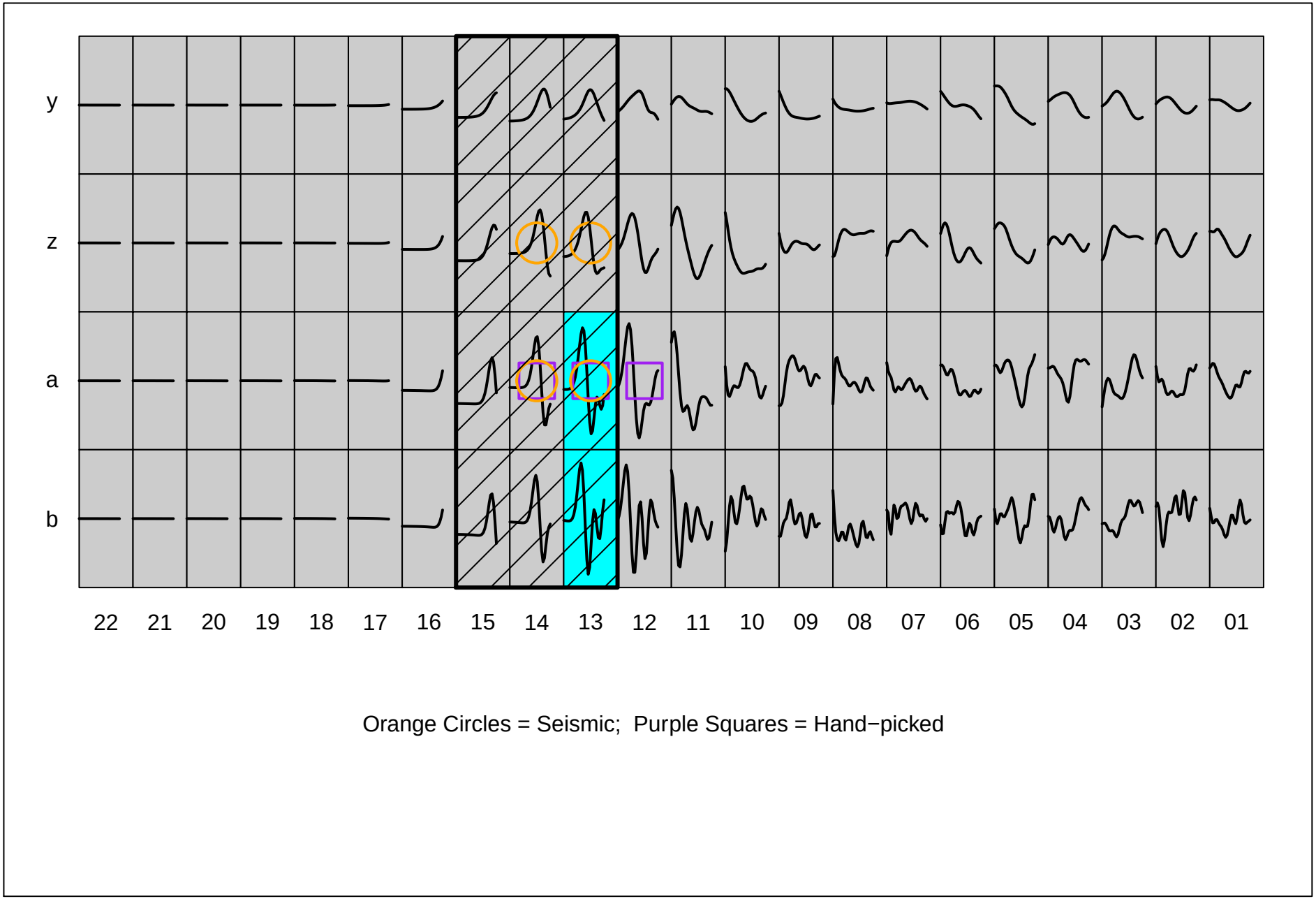
# Buoy 21414: AIC Score −227.3659



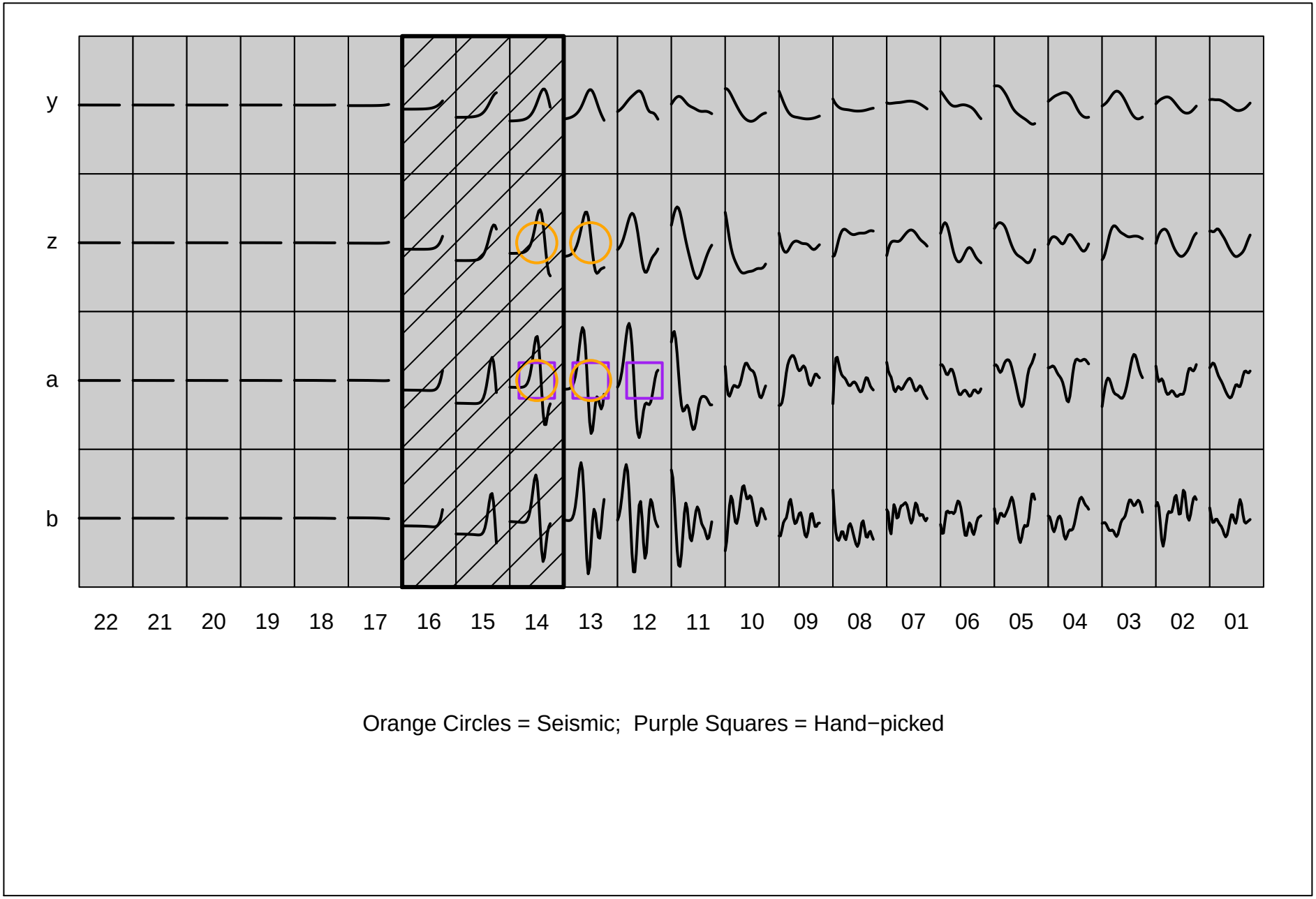
# Buoy 21414: AIC Score −229.101



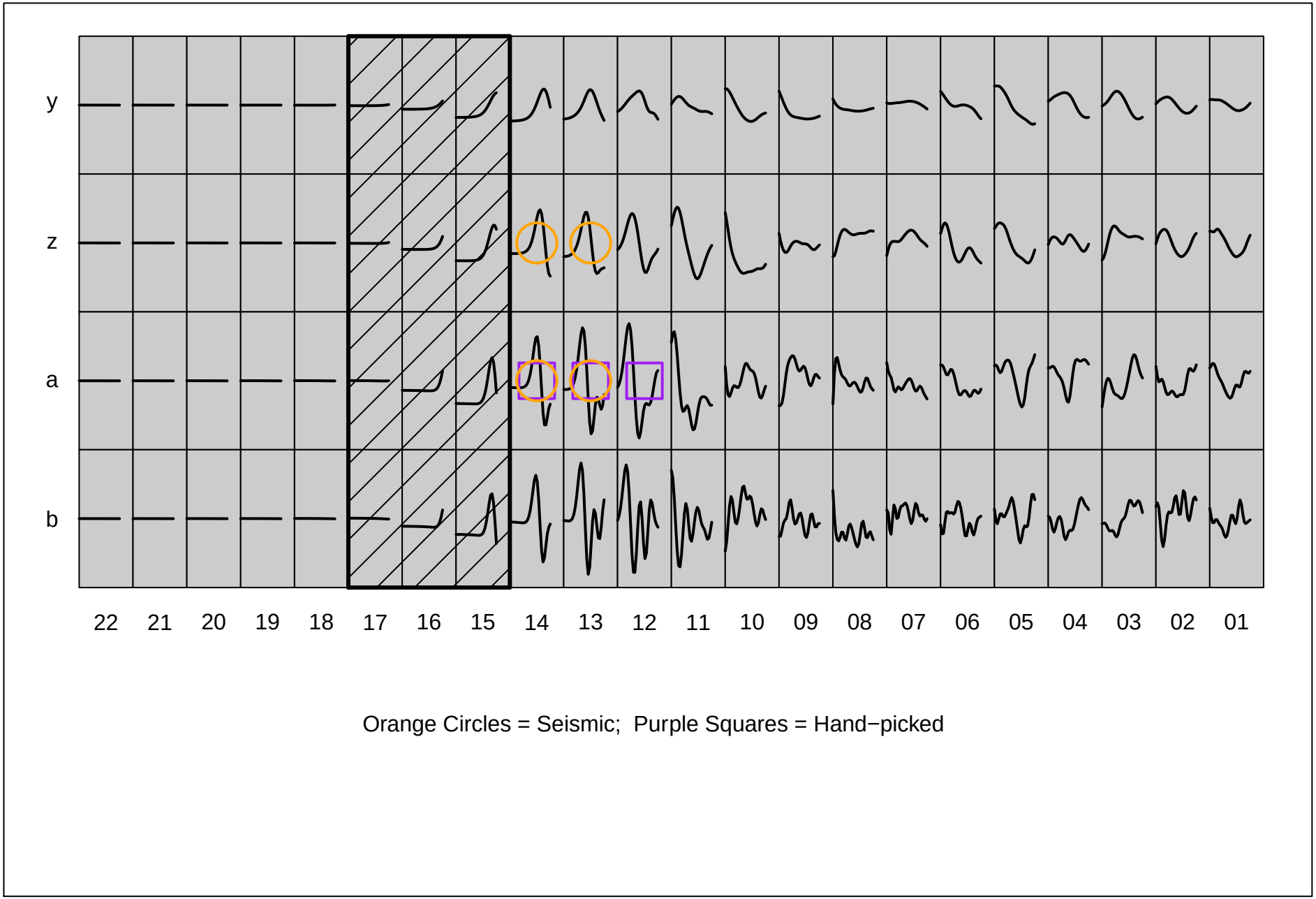
# Buoy 21414: AIC Score −127.0276



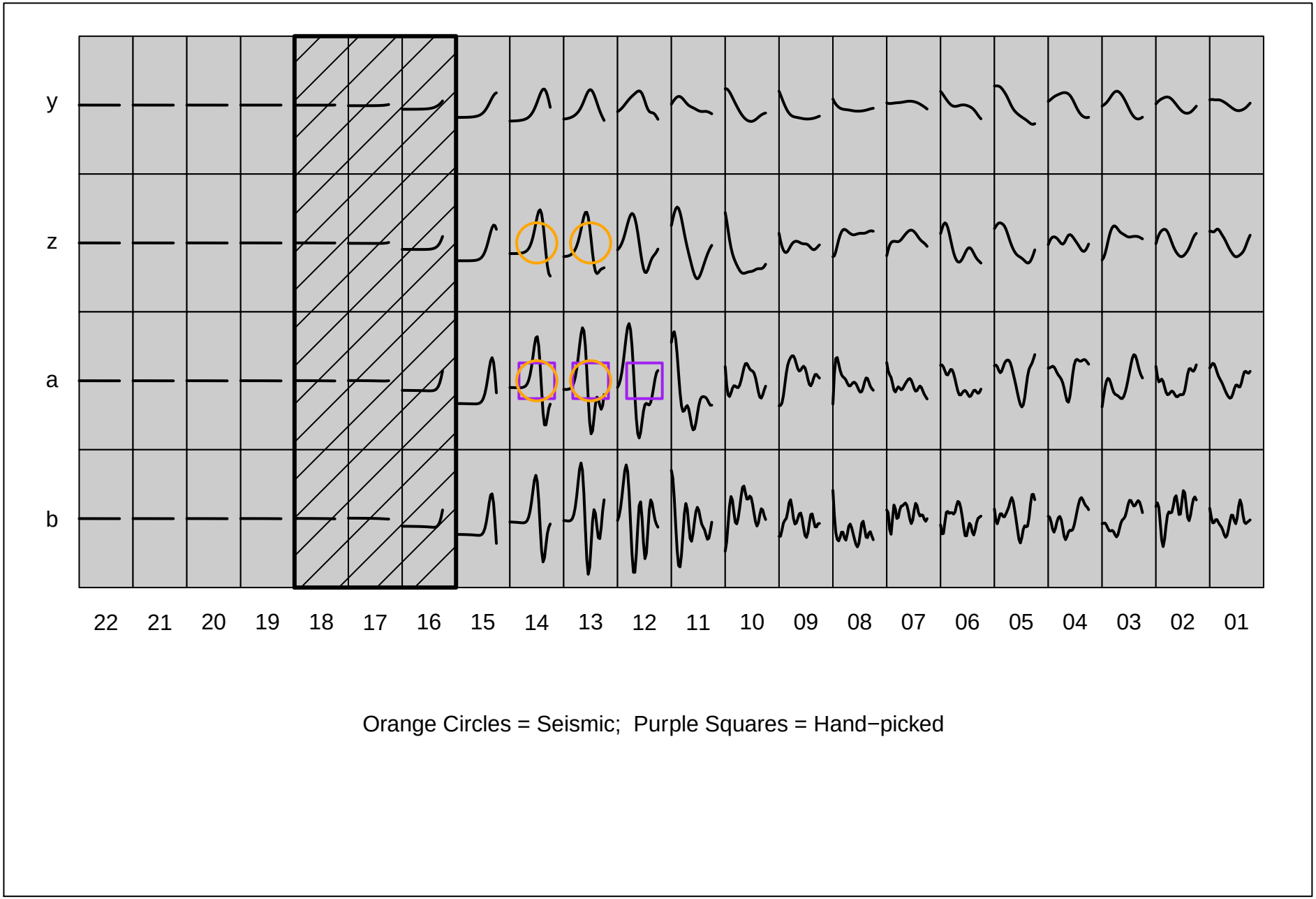
# Buoy 21414: AIC Score −122.8054



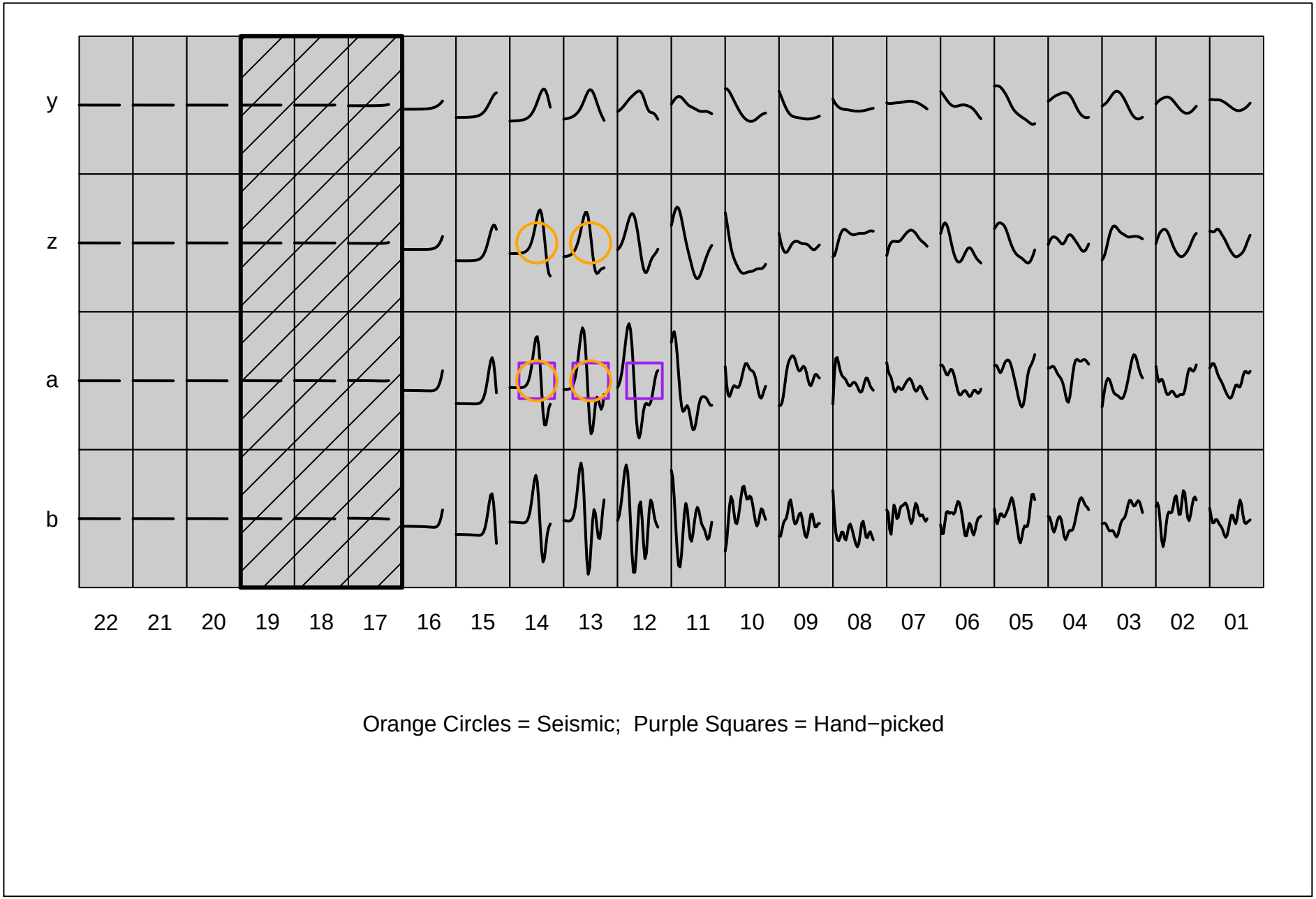
# Buoy 21414: AIC Score −122.8054



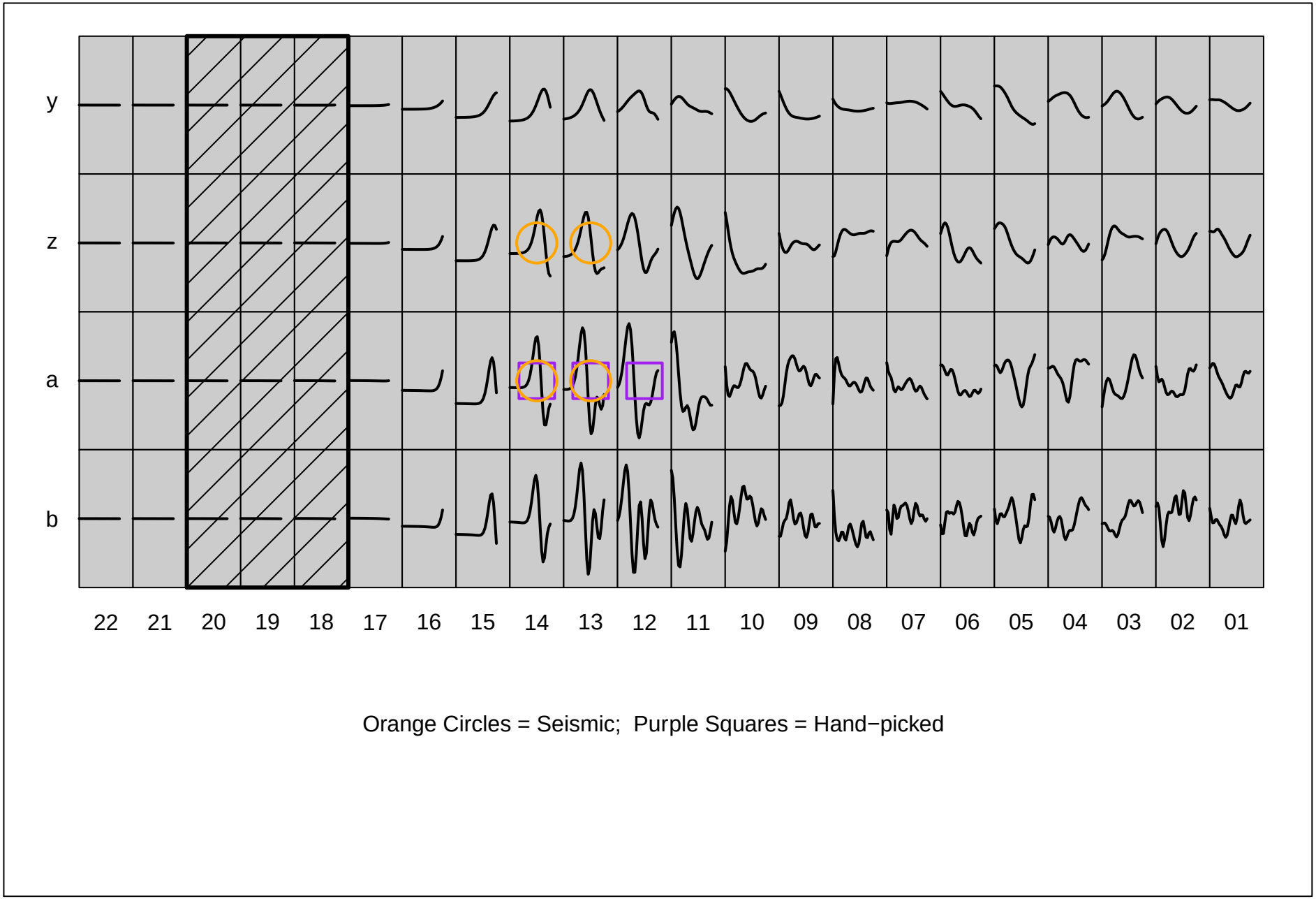
# Buoy 21414: AIC Score −122.8054



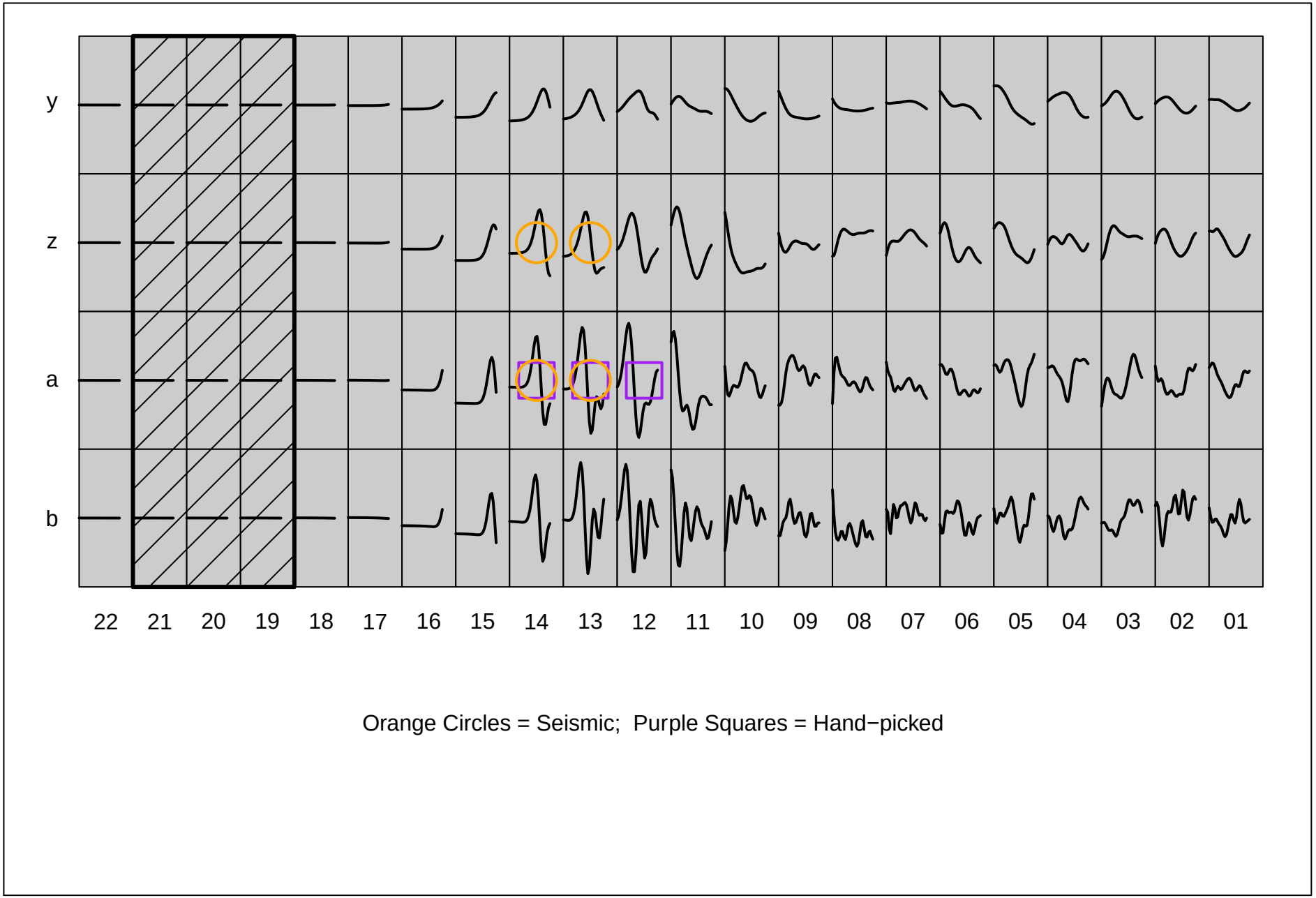
# Buoy 21414: AIC Score −122.8054



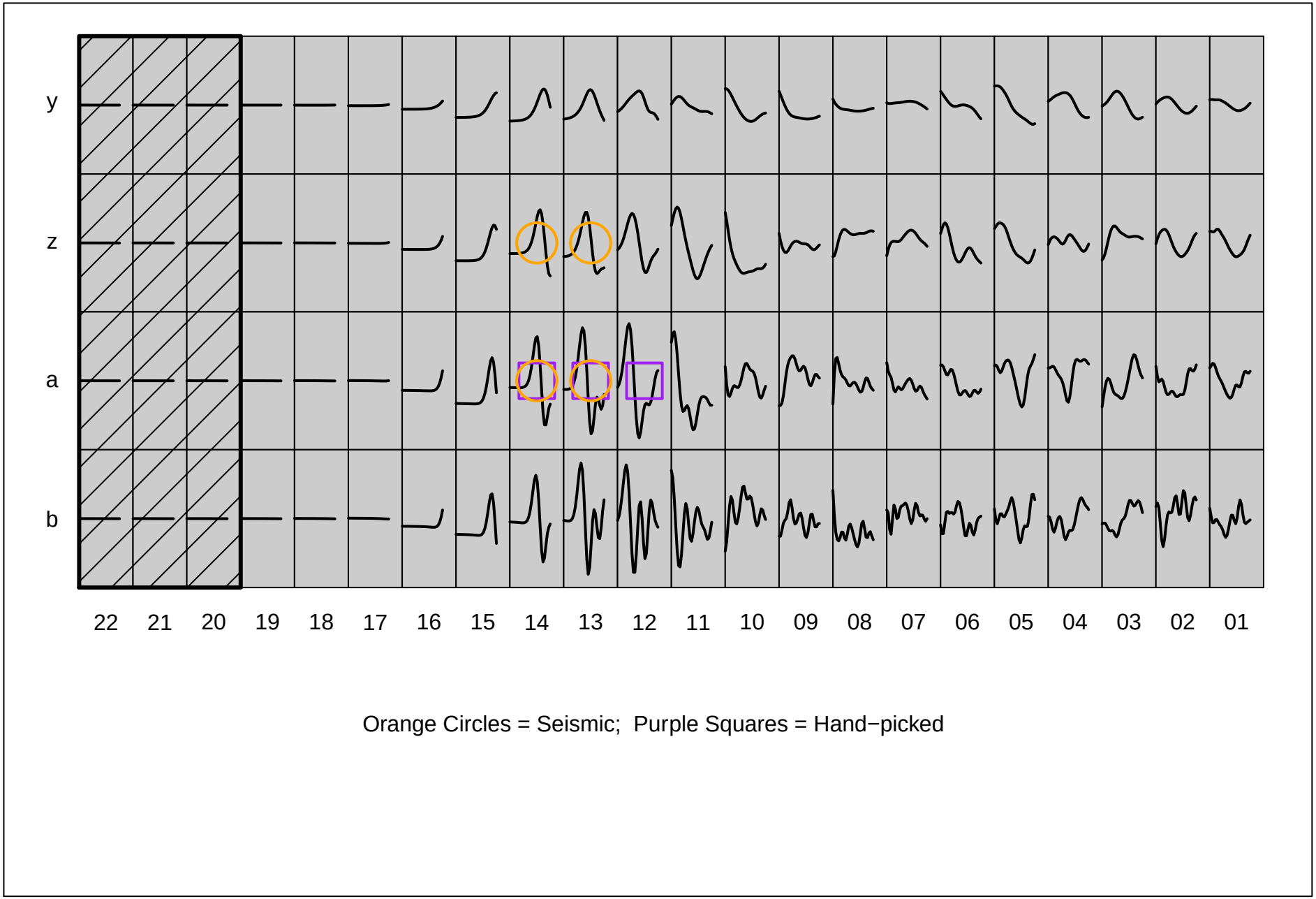
# Buoy 21414: AIC Score −122.8054



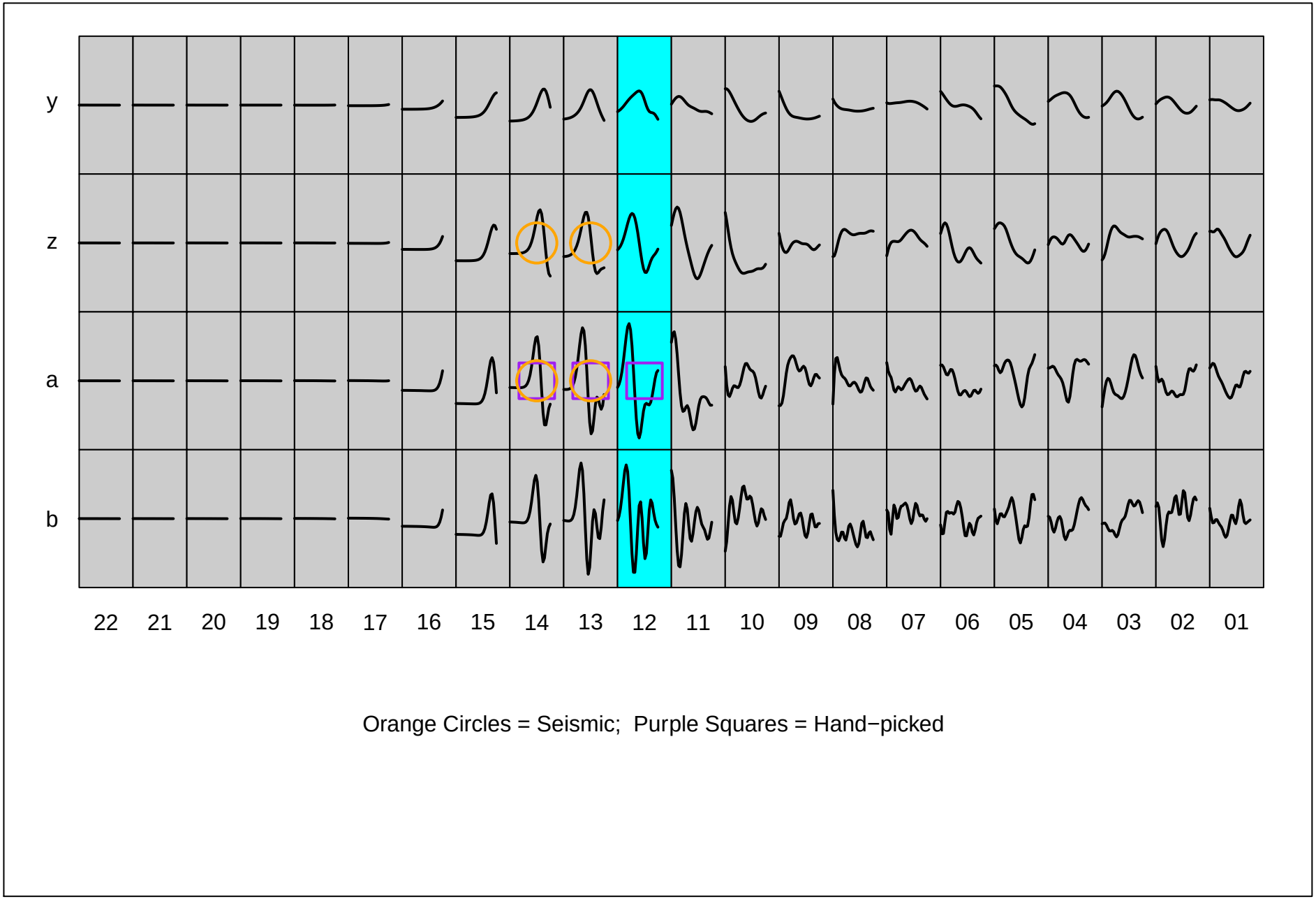
# Buoy 21414: AIC Score −122.8054



# Buoy 21414: AIC Score −122.8054



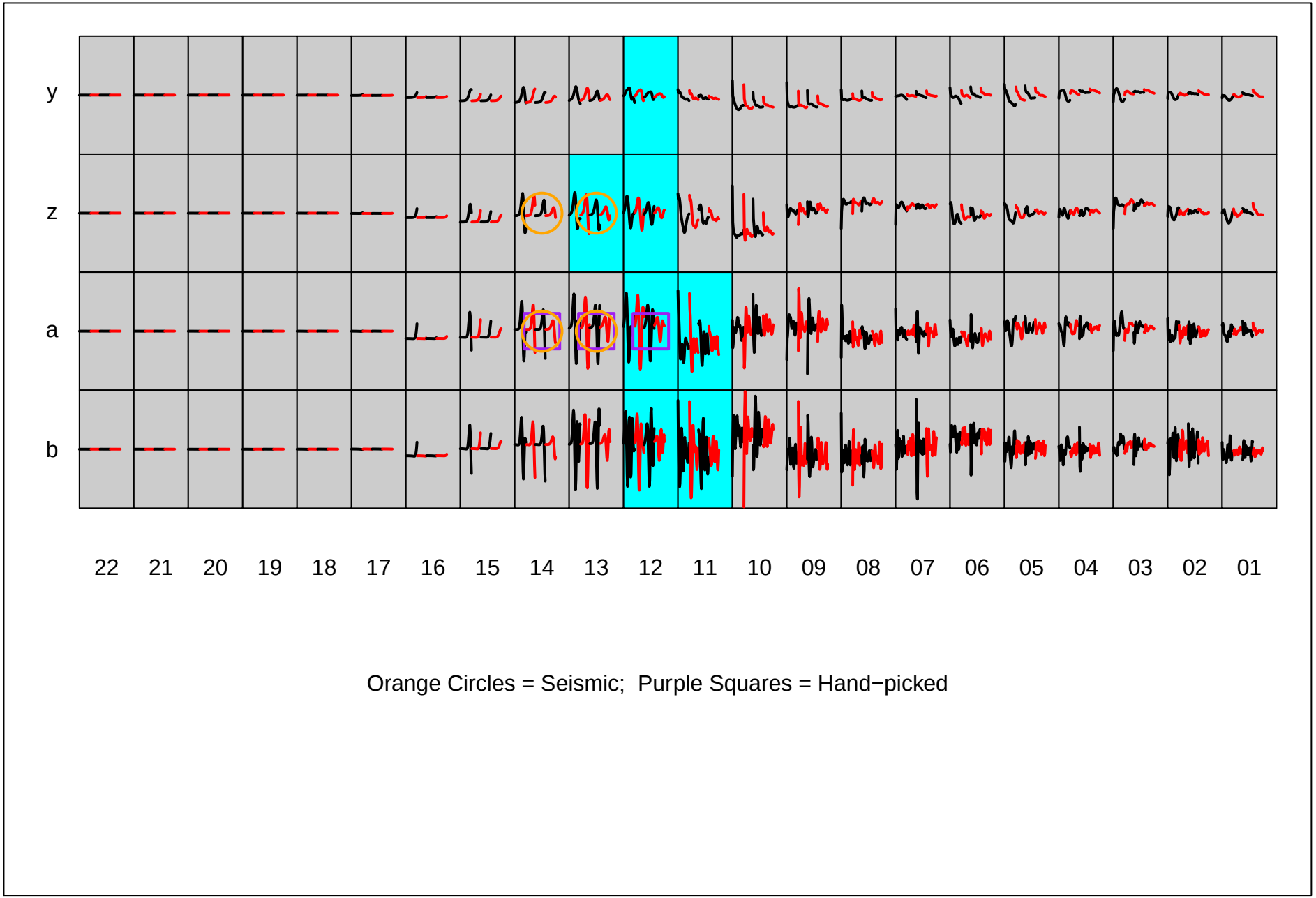
# Buoy 21414: Sweeping Result chosen with AIC



## Selection of Predictors and $\alpha$ Estimation: VIII

- sweeping lasso with AIC selects column of 4 unit sources when using just buoy 21414
- sweeping lasso with AIC selects 7 unit sources when using 4 buoys (21414 plus 46413, 46408 and 46402)

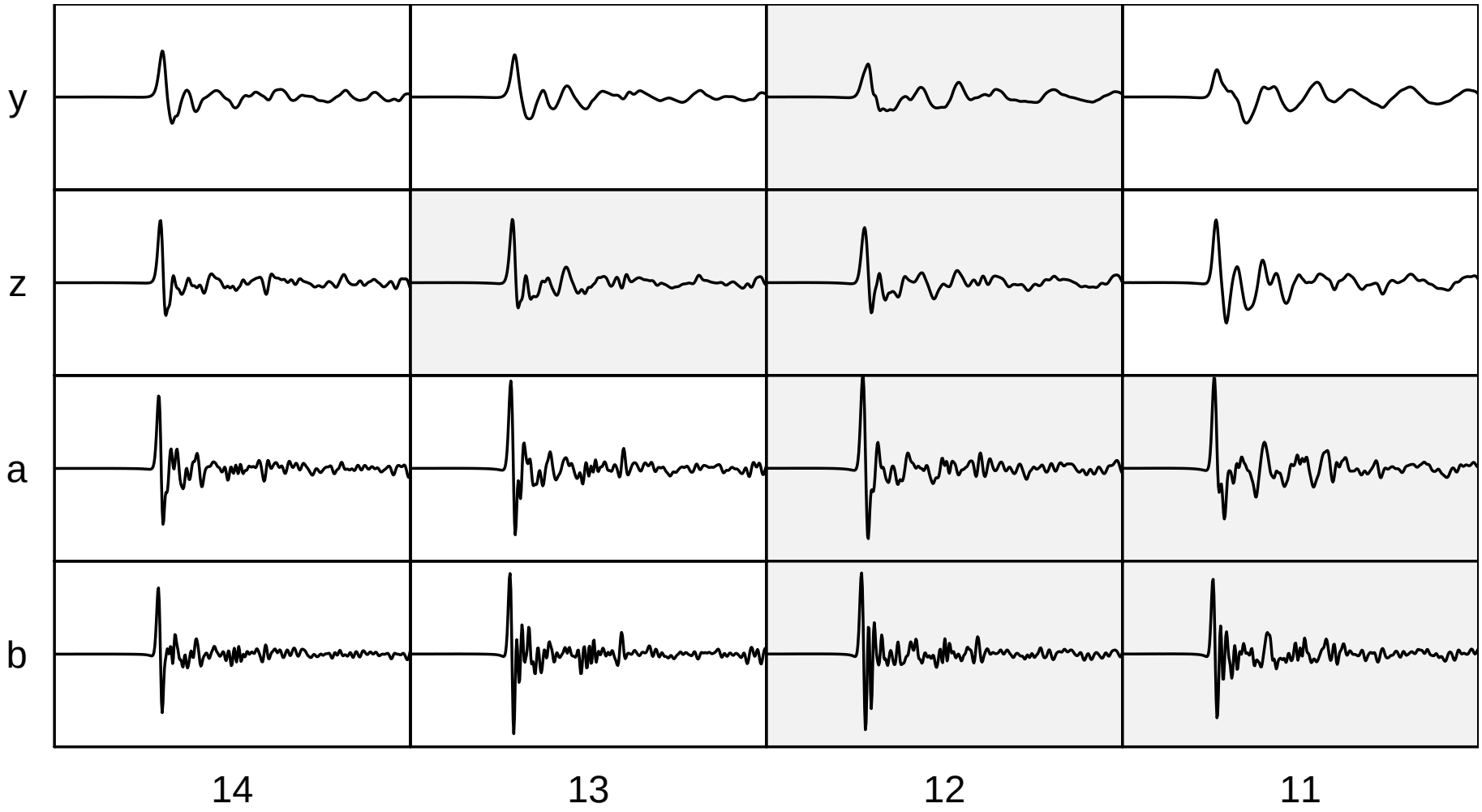
# Buoys 21414, 46413, 46408, 46402, phi = 0.7: Unit Sources Selected by Lasso



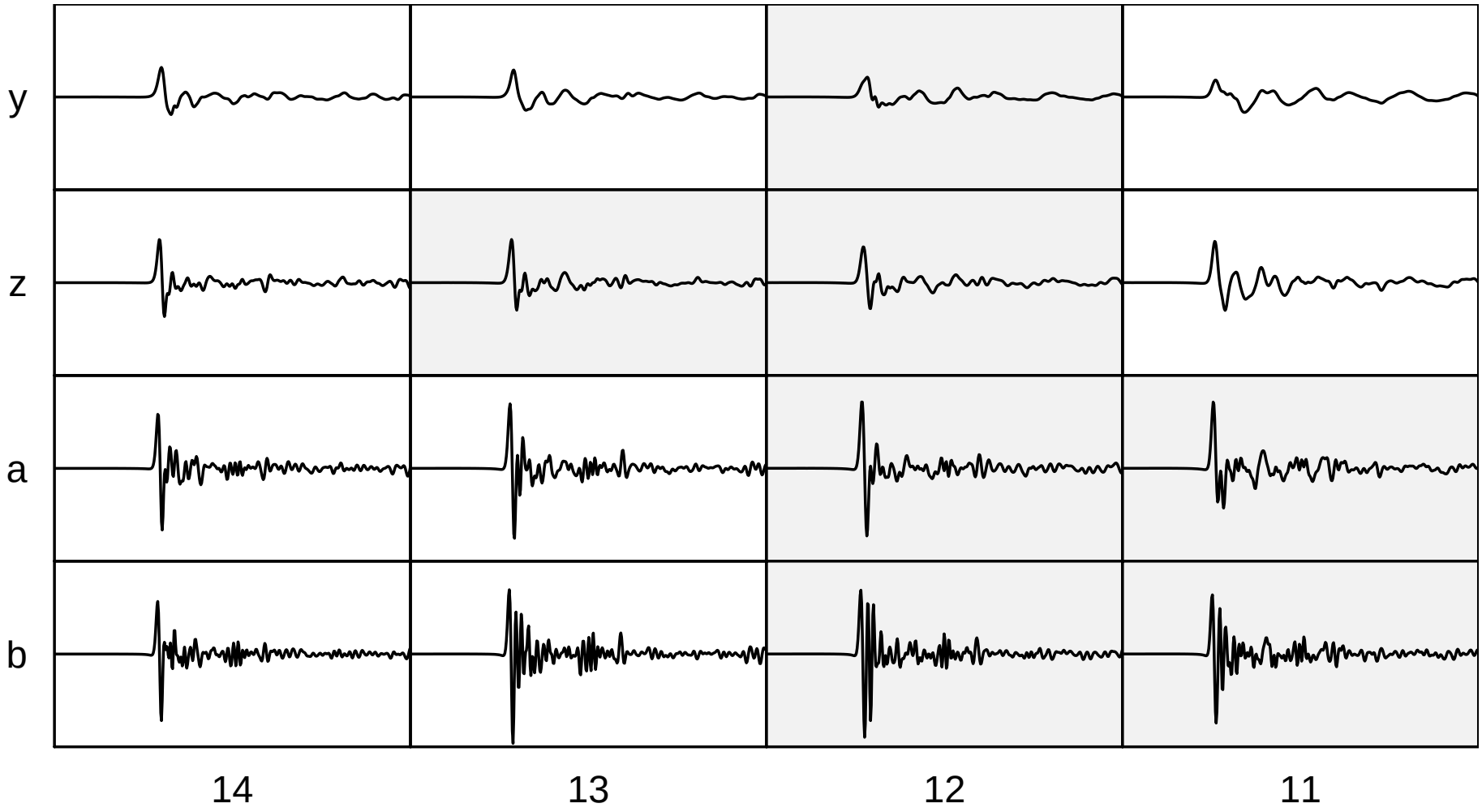
# Collinearity and Ridge Regression: I

- concern with sweeping lasso is collinearity, i.e., high degree of cross-correlation between predictors

# Predictors for Buoy 21414 (0 to 8 hours)



# Transformed Predictors for Buoy 21414 ($\phi = 0.7$)



# Collinearity and Ridge Regression: II

- maximum cross-correlation between displayed transformed predictors for 21414 is 0.75
- *ridge regression* (Hoerl and Kennard, 1970) designed to handle collinearity
- usual (unconstrained) OLS estimator minimizes

$$\|\tilde{\mathbf{d}} - \widetilde{G}\boldsymbol{\alpha}\|_2^2 \text{ and is given by } \hat{\boldsymbol{\alpha}} = (\widetilde{G}^T\widetilde{G})^{-1}\widetilde{G}^T\tilde{\mathbf{d}}$$

- ridge regression estimator given by

$$\hat{\boldsymbol{\alpha}} = (\widetilde{G}^T\widetilde{G} + \lambda_2 I)^{-1}\widetilde{G}^T\tilde{\mathbf{d}},$$

where $\lambda_2$ is a parameter to be set

# Elastic Net: I

- ridge regression is solution to following problem: minimize

$$\|\tilde{\mathbf{d}} - \widetilde{G}\boldsymbol{\alpha}\|_2^2 + \lambda_2\|\boldsymbol{\alpha}\|_2^2,$$

  where $\lambda_2 \geq 0$, and $\lambda_2\|\boldsymbol{\alpha}\|_2$ is the $\ell_2$ penalty

- *elastic net* (Zou and Hastie, 2005) is solution to following problem: minimize

$$\|\tilde{\mathbf{d}} - \widetilde{G}\boldsymbol{\alpha}\|_2^2 + \lambda_1\|\boldsymbol{\alpha}\|_1 + \lambda_2\|\boldsymbol{\alpha}\|_2^2$$

- elastic net is lasso with $\ell_2$ penalty added to tackle collinearity

- note: need to modify standard elastic net to include nonnegativity constraint $\boldsymbol{\alpha} \geq \mathbf{0}$

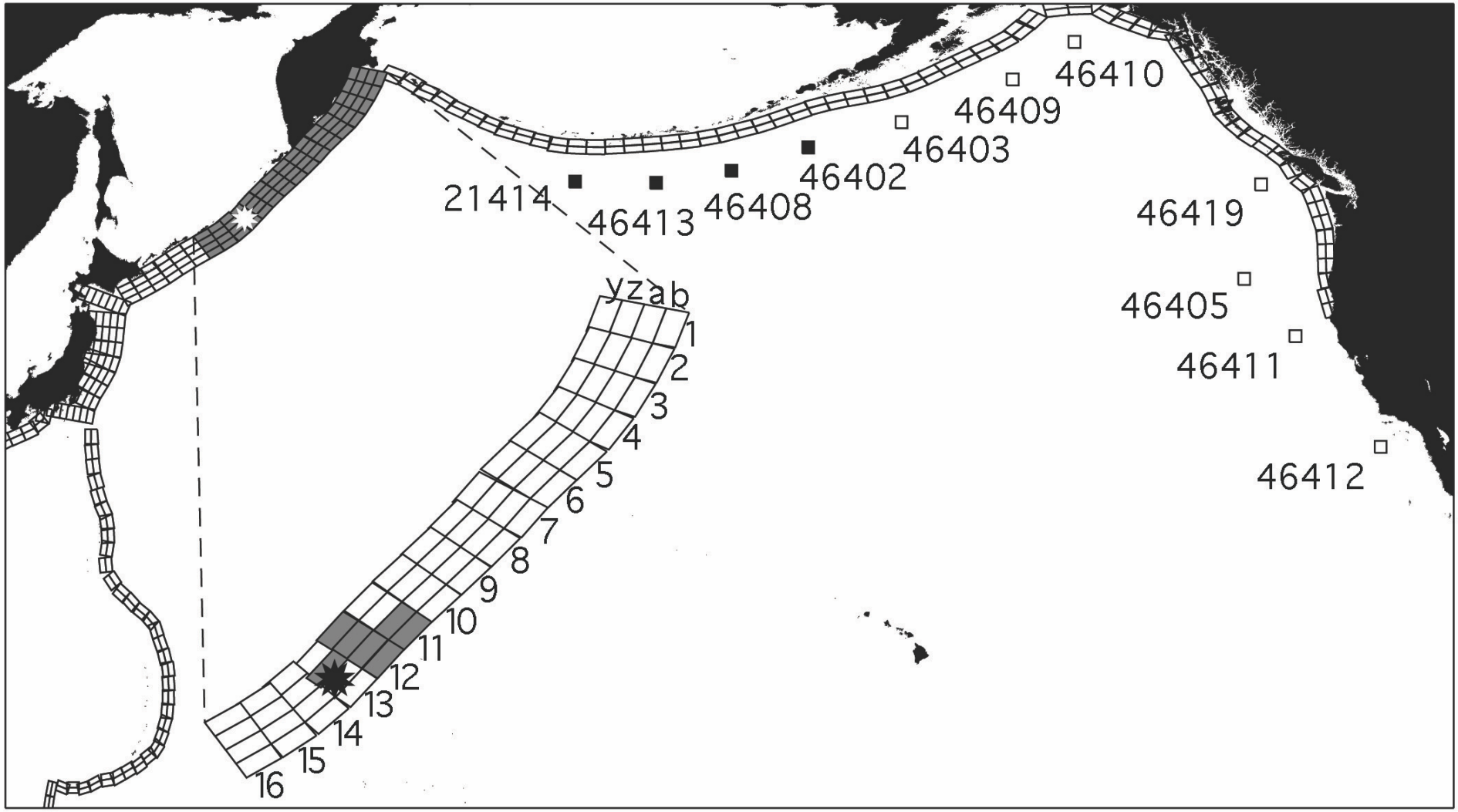
# Elastic Net: II

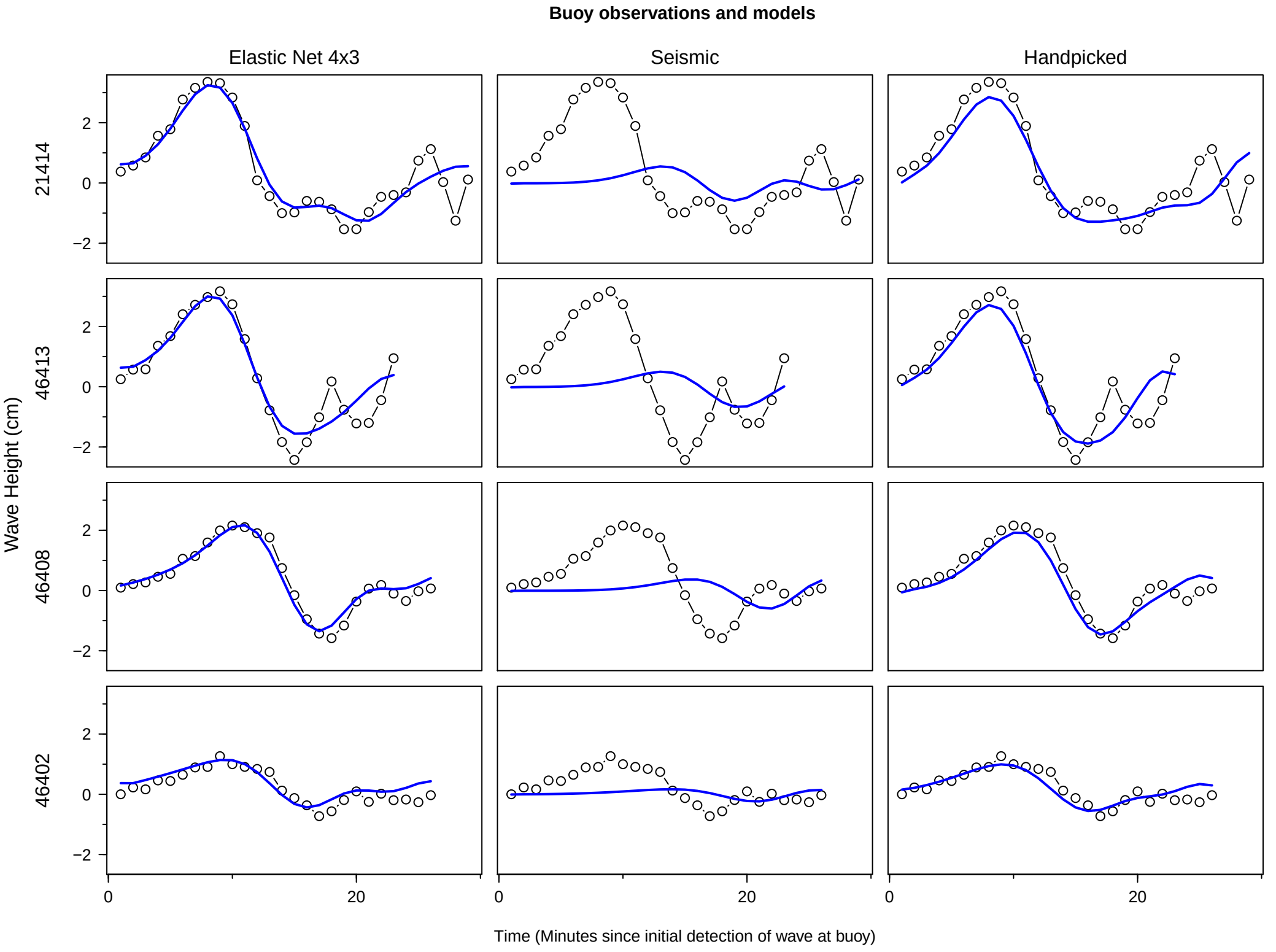
- without the $\ell_2$ penalty
  - high variance in $\hat{\boldsymbol{\alpha}}$, increasing chance of inaccurate magnitude prediction
  - signs in $\hat{\boldsymbol{\alpha}}$ can flip, resulting in exclusion in constrained estimation
- with the $\ell_2$ penalty
  - variance of estimators $\hat{\boldsymbol{\alpha}}$ stabilized
  - groups of correlated predictors can enter together

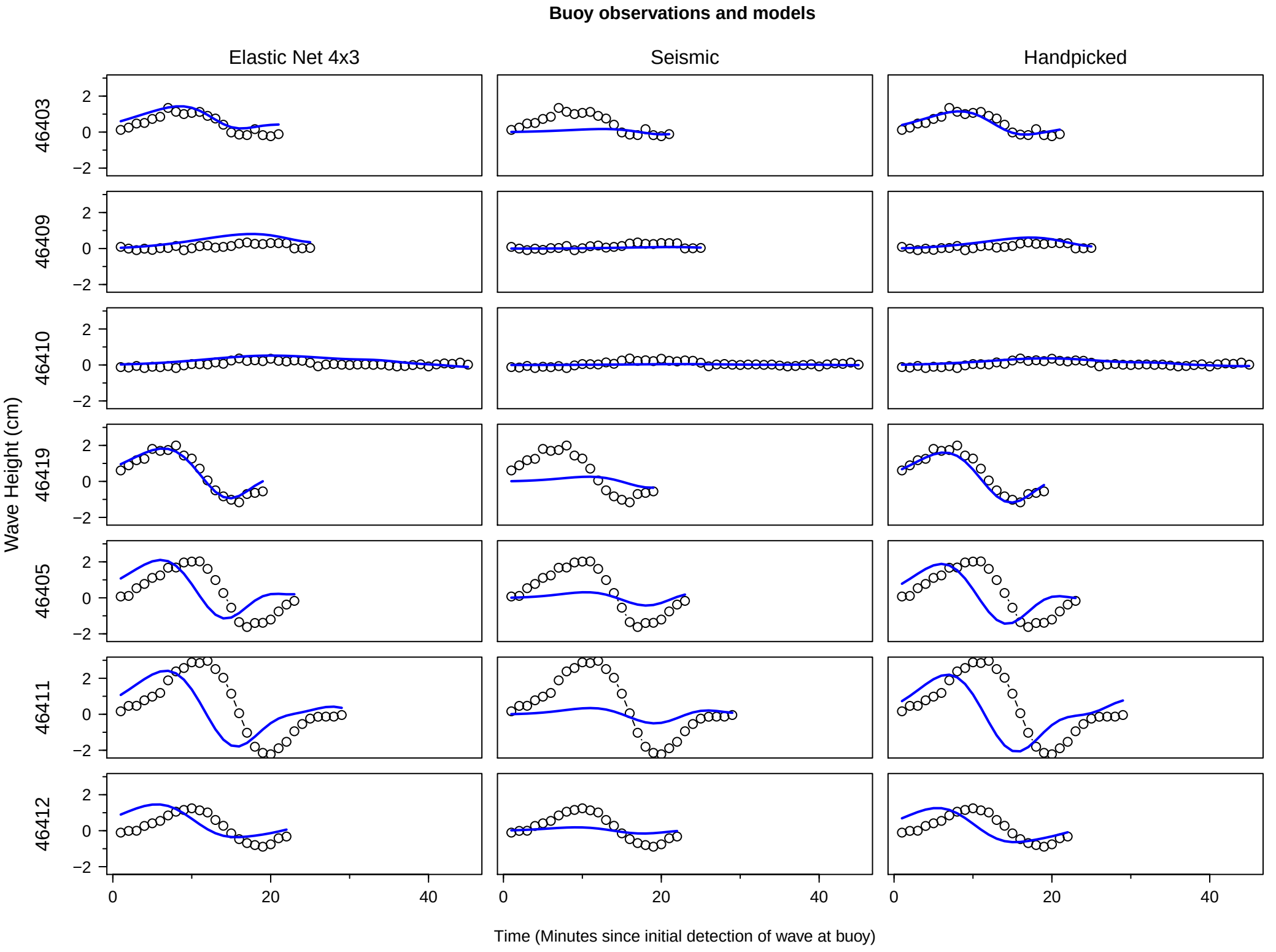
# Elastic Net: III

- elastic net has two tuning parameters: $\lambda_1$ and $\lambda_2$
- can adjust AIC to select both parameters and to settle on best localized region in sweeping
- experience to date suggests that solutions are robust to different choices of $\lambda_2$ except for values too close to zero or too large
- can set $\lambda_2 = 1$ and then use AIC to handle sweeping elastic net in a manner similar to sweeping lasso
- for Kuril Islands example, sweeping elastic net yields same choice of predictors as sweeping lasso
  - in fact, sweeping ridge regression yields identical results!

# Predictors for Nov. 2006 Kuril Islands Event



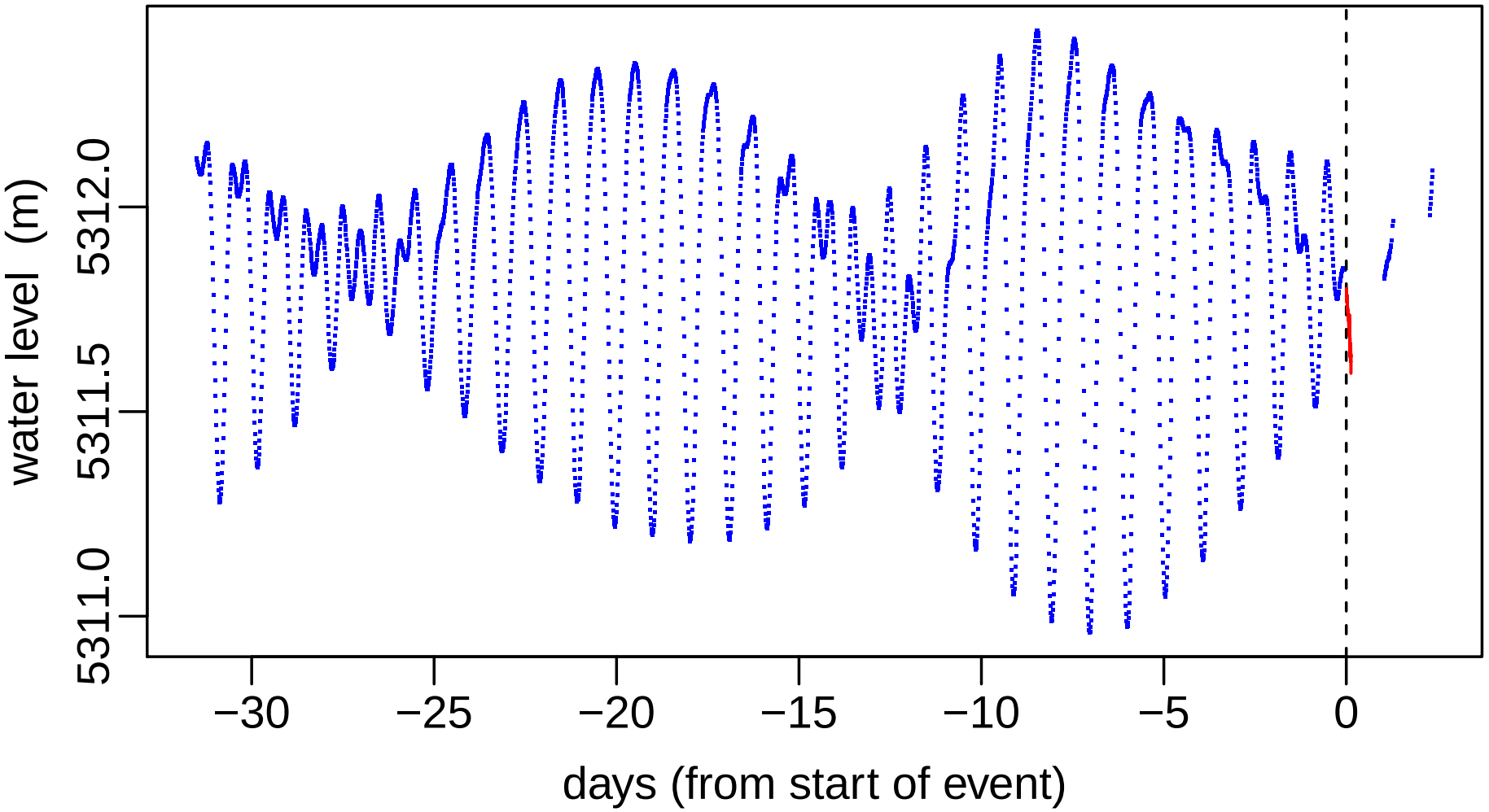

Buoy observations and models
Elastic Net 4x3
Seismic
Handpicked
21414
46413
46408
46402
Wave Height (cm)
2
0
−2
0
20
Time (Minutes since initial detection of wave at buoy)


Buoy observations and models
Elastic Net 4x3
Seismic
Handpicked
46403
46409
46410
46419
46405
46411
46412
Wave Height (cm)
2
0
−2
0
20
40
Time (Minutes since initial detection of wave at buoy)

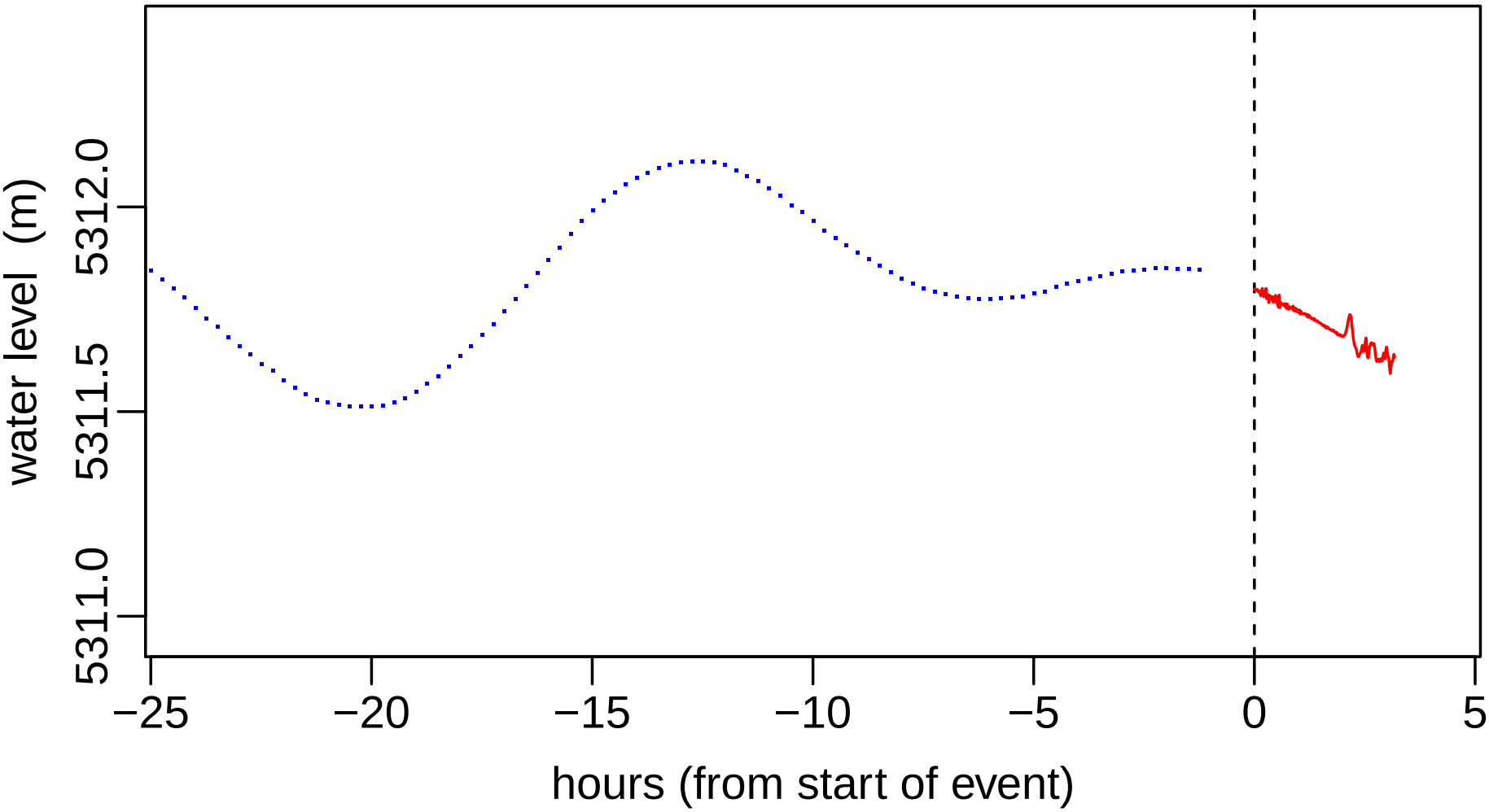
# Accounting for Tides and Background Noise: I

- predictors based on geophysical models do not take into account tidal effects, which dominate data collected by DART® buoys

# Buoy 21414 Data for Nov. 2006 Kuril Islands Event



# Buoy 21414 Data for Nov. 2006 Kuril Islands Event



# Accounting for Tides and Background Noise: II

- accounting for tides and background noise not a trivial task
- operational model for $j$th buoy's data: $\bar{\mathbf{y}}_j = \mathbf{x}_j + G_j\boldsymbol{\alpha} + \boldsymbol{\epsilon}_j$, which will be simplified by dropping subscripts: $\bar{\mathbf{y}} = \mathbf{x}+G\boldsymbol{\alpha}+\boldsymbol{\epsilon}$
  - $\bar{\mathbf{y}}$ is vector with 1-min stream; i.e., averages of four consecutive 15-sec bottom pressure (BP) measurements from DART® buoy available during a tsunami event
  - $\mathbf{x}$ is vector representing tidal fluctuations
  - $G$ is a matrix whose $K$ columns contain predictors that collectively model tsunami signal
  - $\boldsymbol{\alpha} \geq \mathbf{0}$ is a vector of source coefficients
  - $\boldsymbol{\epsilon}$ is vector representing background noise

# Accounting for Tides and Background Noise: III

- have studied five approaches to detiding (*many* other approaches!)
  1. harmonic analysis based on 29 days of data prior to event
  2. harmonic analysis based on lots of prior data (300–1000 days)
  3. empirical orthogonal function (EOF) approach
  4. Kalman smoothing (KS) approach
  5. harmonic analysis with joint estimation of source coefficients
- first two methods are based on physical models
- first four methods predict tidal fluctuations using, say, $\hat{\mathbf{x}}$
- predictions are subtracted from $\bar{\mathbf{y}}$ to form detided data:

$$\mathbf{d} = \bar{\mathbf{y}} - \hat{\mathbf{x}} = G\boldsymbol{\alpha} + \mathbf{e},$$

where $\mathbf{e} = \boldsymbol{\epsilon} + \mathbf{x} - \hat{\mathbf{x}}$ is error term (includes background noise and inaccuracies in predicting tidal fluctuations)

# Accounting for Tides and Background Noise: IV

- can use $\mathbf{d}$ from any of the first four methods with sweeping elastic net (Kuril Islands example based on KS approach)
- fifth method handles tidal fluctuations and estimation of source coefficients $\boldsymbol{\alpha}$ jointly least ('joint method')
- in contrast to previous four methods, joint method does *not* make use of any data prior to tsunami event – just uses available 1-min stream
  - simplifies matters operationally – part or all of 15-min stream has been missing in past events
- will now briefly describe five methods

# Method 1 (29 Day Harmonic Analysis)

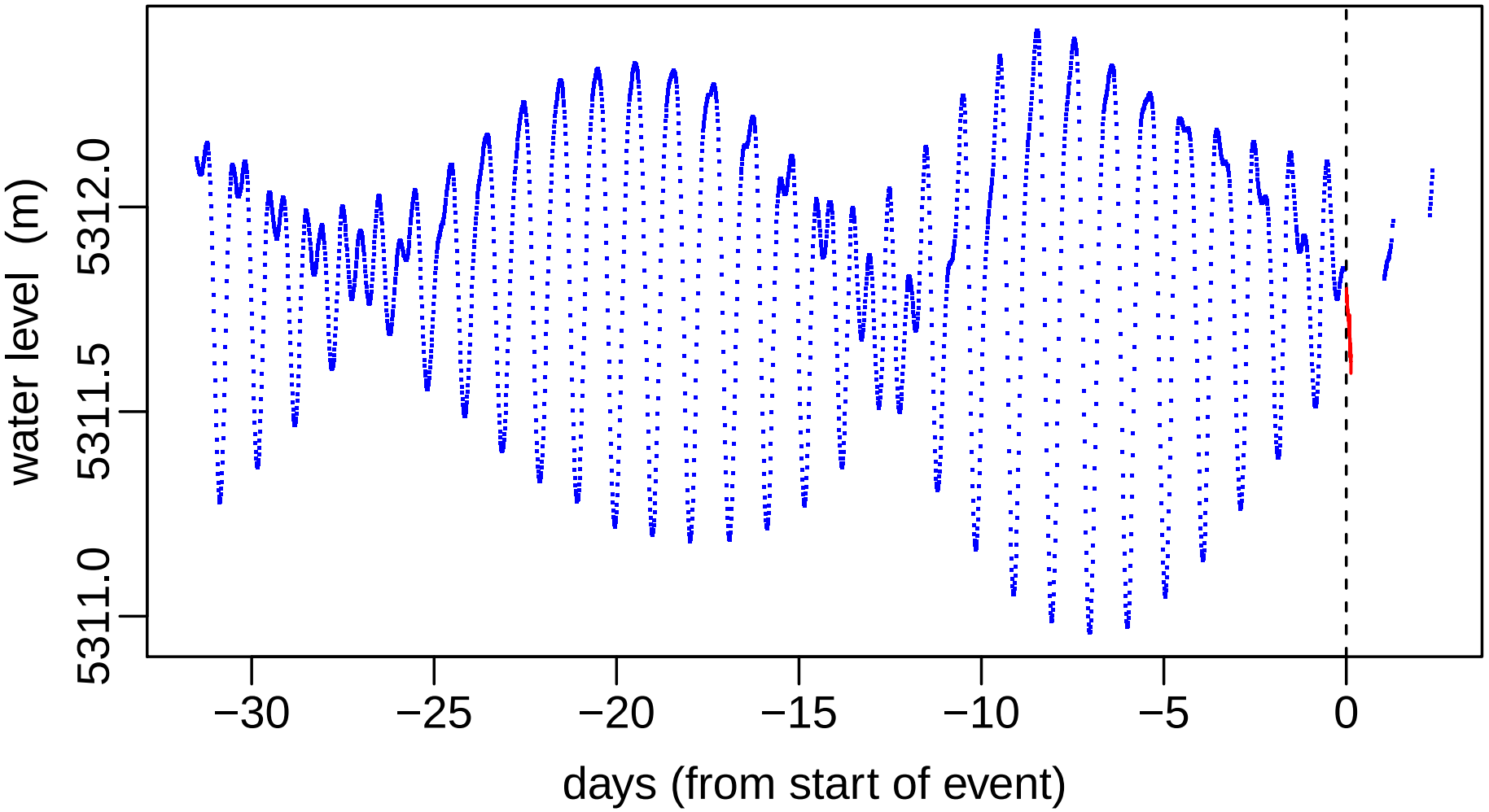
- harmonic analysis is standard way to predict tides at coastal stations and assumes tides are sums of sinusoidal constituents
- for detiding DART® data, assume buoy has reported BP measurements every 15 minutes for 29 days prior to tsunami event
- model measurements $y_n$ from 15-min stream as

$$y_n = \mu + \sum_{m=1}^{6} [B_m \cos(\omega_m n\,\Delta) + C_m \sin(\omega_m n\,\Delta)] + e_n,$$

  where parameters $\mu$, $B_m$ & $C_m$ are estimated via least squares ($\omega_m$'s for N2, M2, S2, Q1, O1 & K1 and $\Delta$ are known)
- use fitted model to form detided 1-min stream:

$$d_n = \bar{y}_n - \hat{\mu} - \frac{1}{4}\sum_{k=0}^{3}\sum_{m=1}^{6}\left[\widehat{B}_m \cos(\omega_m[n+k]\,\Delta) + \widehat{C}_m \sin(\omega_m[n+k]\,\Delta)\right]$$

# Buoy 21414 Data for Nov. 2006 Kuril Islands Event



# Method 2 (Long Harmonic Analysis)

- similar to method 1, but with following key differences
  - model now uses 68 constituents rather than just 6
  - except for mean level $\mu$, data used to fit model are from 15-sec streams retrieved from buoy during regular servicing (totality of streams typically spans from 300 to 1000 days)
  - 29 days from 15-min stream now used just to set $\mu$
- fitting model requires specialized software (G. Mungov, NOAA National Centers for Environmental Information, provided fits for study discussed later on)
- cannot use method unless buoy has been serviced once in its present location

# Linear Time-Invariant (LTI) Filtering

- two best known approaches for detiding are harmonic analysis and LTI high-pass filtering (e.g., Butterworth filtering)
- to isolate tsunami signal without distortion, filter should retain components with periods as long as 2 hours
- potential disadvantages
  - edge effects can significantly distort at least 1-hr sections at beginning and end of filtered series, rendering approach problematic in real-time environment
  - most LTI filters not designed to work with gappy data
- methods 3 and 4 are linear (but not LTI) filters designed to overcome these disadvantages

## Method 3 (Empirical Orthogonal Functions)

- premise (Tolkova, 2010): sub-space spanned by leading empirical orthogonal functions (EOFs) of tidally dominated data segments same across all DART® buoys
- EOFs obtained from 250 segments (each spanning one lunar-day) from DART® buoy 46412 in 2007
- available buoy data from 1-min and 15-min streams projected against EOFs $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_7$ associated with seven largest eigenvalues (along with constant vector $\mathbf{f}_0$) to obtain coefficients $c_0$, $c_1, \ldots, c_7$
- inverse projection using $c_0$, $c_1$, $\ldots$, $c_7$ yields predicted tides, which are subtracted from buoy data to yield detided data

# Method 4 (Kalman Smoothing)

- Kalman smoothing (KS) widely used to 'optimally' smooth a time series, but optimality depends upon adequate model for underlying dynamics
- KS approach here is two-stage procedure
  1. use 29 day harmonic analysis (method 1) to obtain first-stage detided series, say, $d_n$
  2. use $d_n$ as input to KS based upon local level model (also known as 'random walk plus noise' model') – output from smoother intended to track any tidal component/background noise left over from first-stage detiding
- local level model depends upon just two parameters and estimate of initial state of underlying dynamical system

# Method 5 (Joint Method): I

- joint method estimates coefficients for tidal model along with $\boldsymbol{\alpha}$ using just 1-min stream transmitted during tsunami event
- joint method is based on operational model

$$\bar{\mathbf{y}} = \mathbf{x} + G\boldsymbol{\alpha} + \boldsymbol{\epsilon}$$

but with addition of specific model for tidal fluctuations:

$$\mathbf{x} = \mu\mathbf{1} + \sum_{m=1}^{M} \left( B_m \mathbf{c}_m + C_m \mathbf{s}_m \right),$$

where $\mathbf{1}$ is a vector of ones; $M$ is either 1 or 2; $\mathbf{c}_m$ is a vector with elements $\cos(\omega_m n\,\Delta)$, where $\omega_1$ is tidal frequency M2, $\omega_2 = 2\omega_1$, $n$ is a time index and $\Delta = 1$ min; and $\mathbf{s}_m$ is analogous to $\mathbf{c}_m$, but with sines replacing cosines

# Method 5 (Joint Method): II

- unknown parameters are source coefficients $\boldsymbol{\alpha}$ and tidal coefficients $\mu$, $B_m$'s and $C_m$'s
- to estimate coefficients, use constrained elastic net: minimize

$$\|\bar{\mathbf{y}} - \mu\mathbf{1} - \sum_{m=1}^{M} (B_m\mathbf{c}_m + C_m\mathbf{s}_m) - G\boldsymbol{\alpha}\|_2^2 + \lambda_1\|\boldsymbol{\alpha}\|_1 + \lambda_2\|\boldsymbol{\alpha}\|_2^2$$

  subject to $\boldsymbol{\alpha} \geq \mathbf{0}$
- if so desired, can take detided series for joint method to be

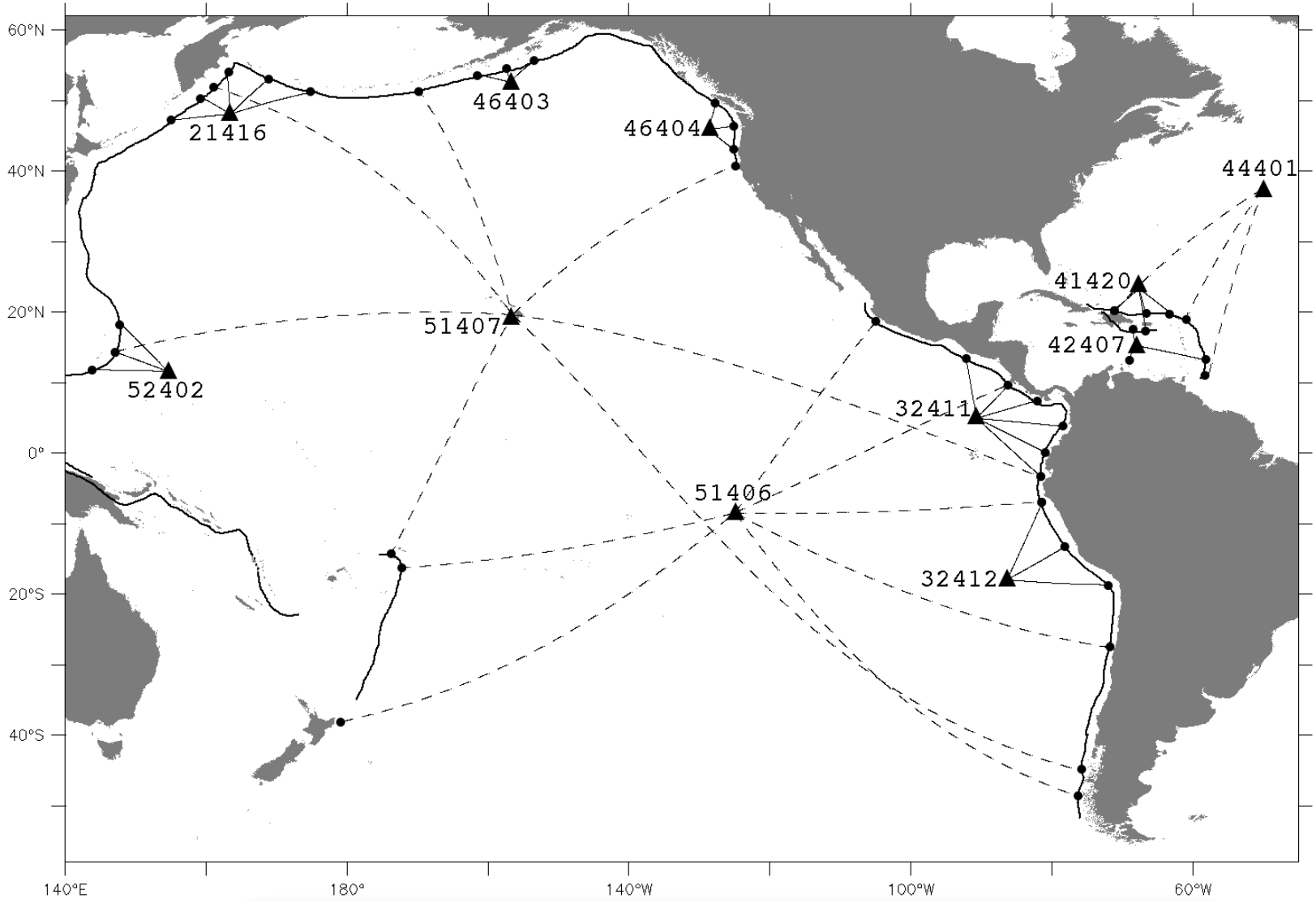
$$\mathbf{d} = \bar{\mathbf{y}} - \hat{\mu}\mathbf{1} - \sum_{m=1}^{M} \left(\widehat{B}_m\mathbf{c}_m + \widehat{C}_m\mathbf{s}_m\right),$$

  where $\hat{\mu}$ is estimate of $\mu$ etc.

# Assessing Performance of Five Methods: I

- recall overall goal: use data from DART® buoys to estimate source coefficients $\boldsymbol{\alpha}$ – these are used in forming coastal inundation forecasts
- Q: how do estimated source coefficients compare for five detiding methods?
- to address question, carried out study using archived 15-sec streams from eleven representative DART® buoys (streams ranged in length from 321 to 998 days)

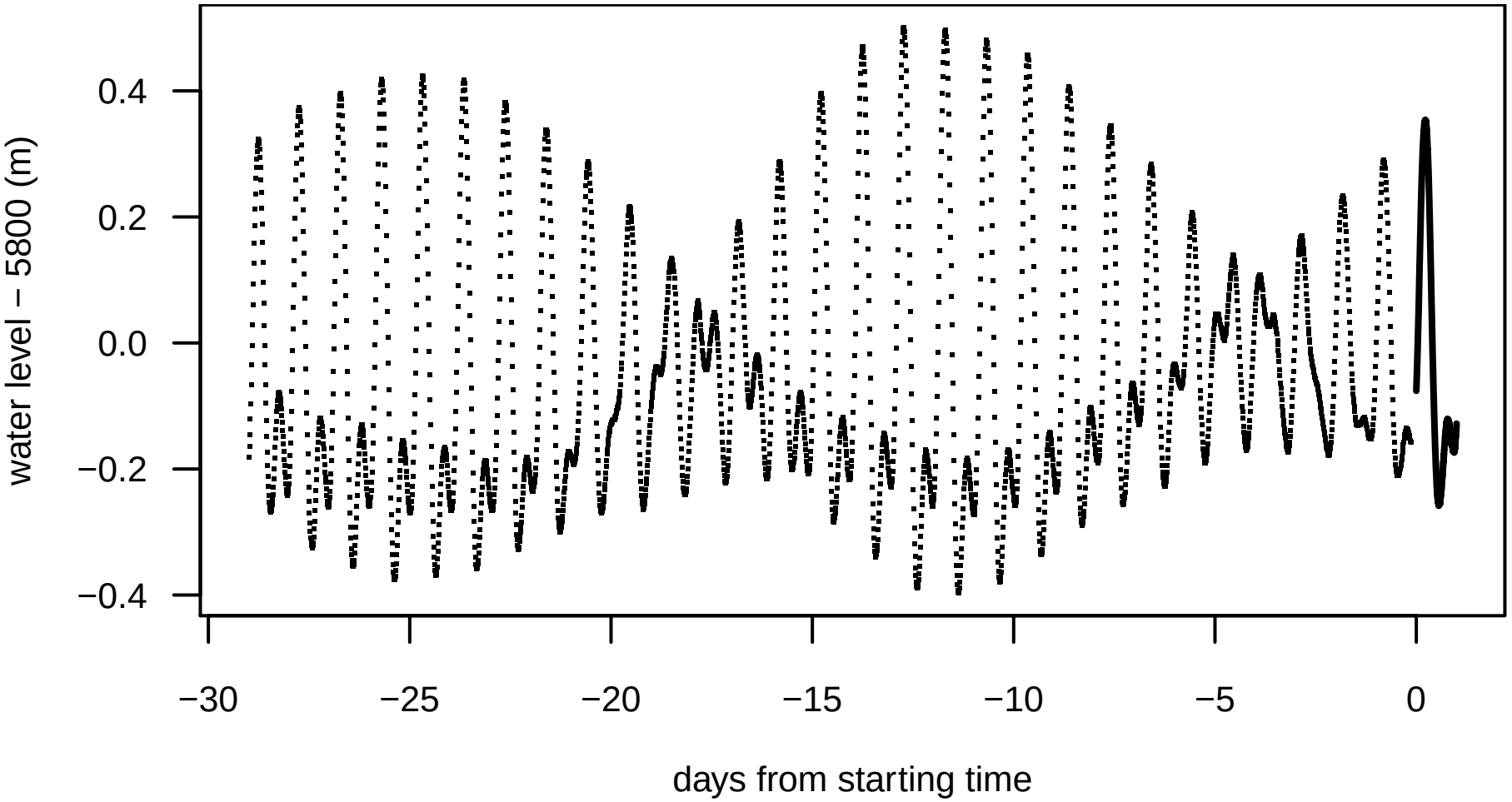
# Locations of Eleven DART® Buoys (Triangles)



# Assessing Performance of Five Methods: II

- recall operational model: $\bar{\mathbf{y}} = \mathbf{x} + G\boldsymbol{\alpha} + \boldsymbol{\epsilon}$
- used archived 15-sec streams to construct 'scenarios' that mimic tidal fluctuations and background noise (i.e., $\mathbf{x} + \boldsymbol{\epsilon}$) present in 15-min & 1-min streams recorded during actual tsunami event
- procedure for constructing one scenario for a particular buoy
  - select random starting time $t_0$
  - form 29-day segment of 15-min stream by subsampling 15-sec stream prior to $t_0$ (to mimic operational conditions, create 3-h gap just prior to $t_0$ in constructed 15-min stream)
  - form 1-day segment of 1-min stream by averaging 4 adjacent values of 15-sec stream after $t_0$
- constructed 15-min & 1-min streams form one of 1000 scenarios

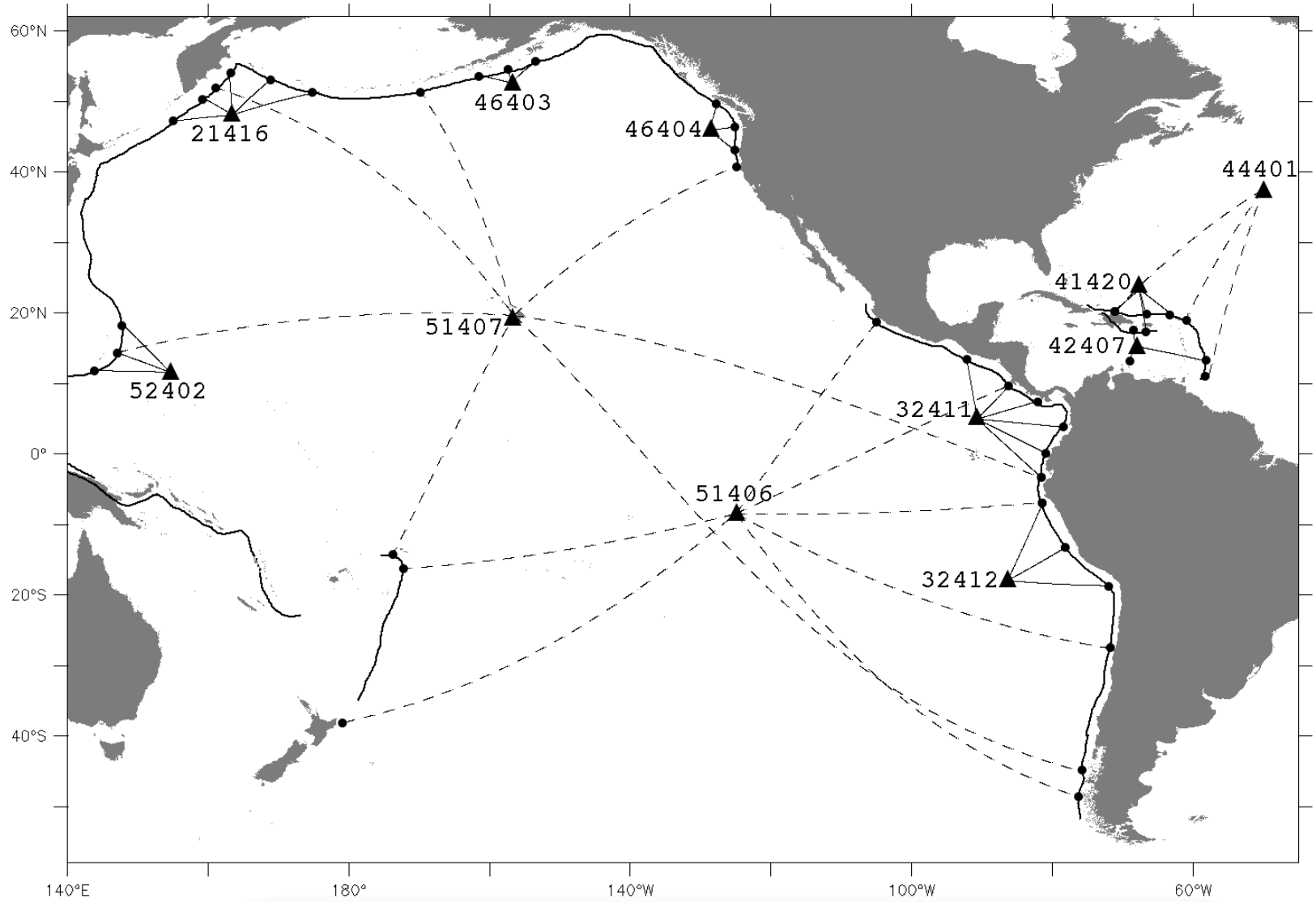
# Scenario 943 for Buoy 52402 ($t_0$ = 9:21:00 UT, 6/27/07)



# Assessing Performance of Five Methods: III

- to create an artificial tsunami signal, simplify operational model to have just one unit source: $\bar{\mathbf{y}} = \mathbf{x} + \alpha\mathbf{g} + \boldsymbol{\epsilon}$
- motivated by actual tsunami events, set $\alpha = 6$ as representative source coefficient
- for each buoy, set $\mathbf{g}$ using three to seven unit sources with different orientations with respect to buoy
  - for example, buoy 52402 is paired with three unit sources (from north to south, ki050b, ki055b and ki060b)
- 42 unit sources in all, five of which were used by two buoys, for a total of 47 pairings of buoys and unit sources (each pairing leads to a different artificial tsunami signal)
- artificial tsunami signal for given buoy/unit source pairing added to each of 1000 scenarios for given buoy

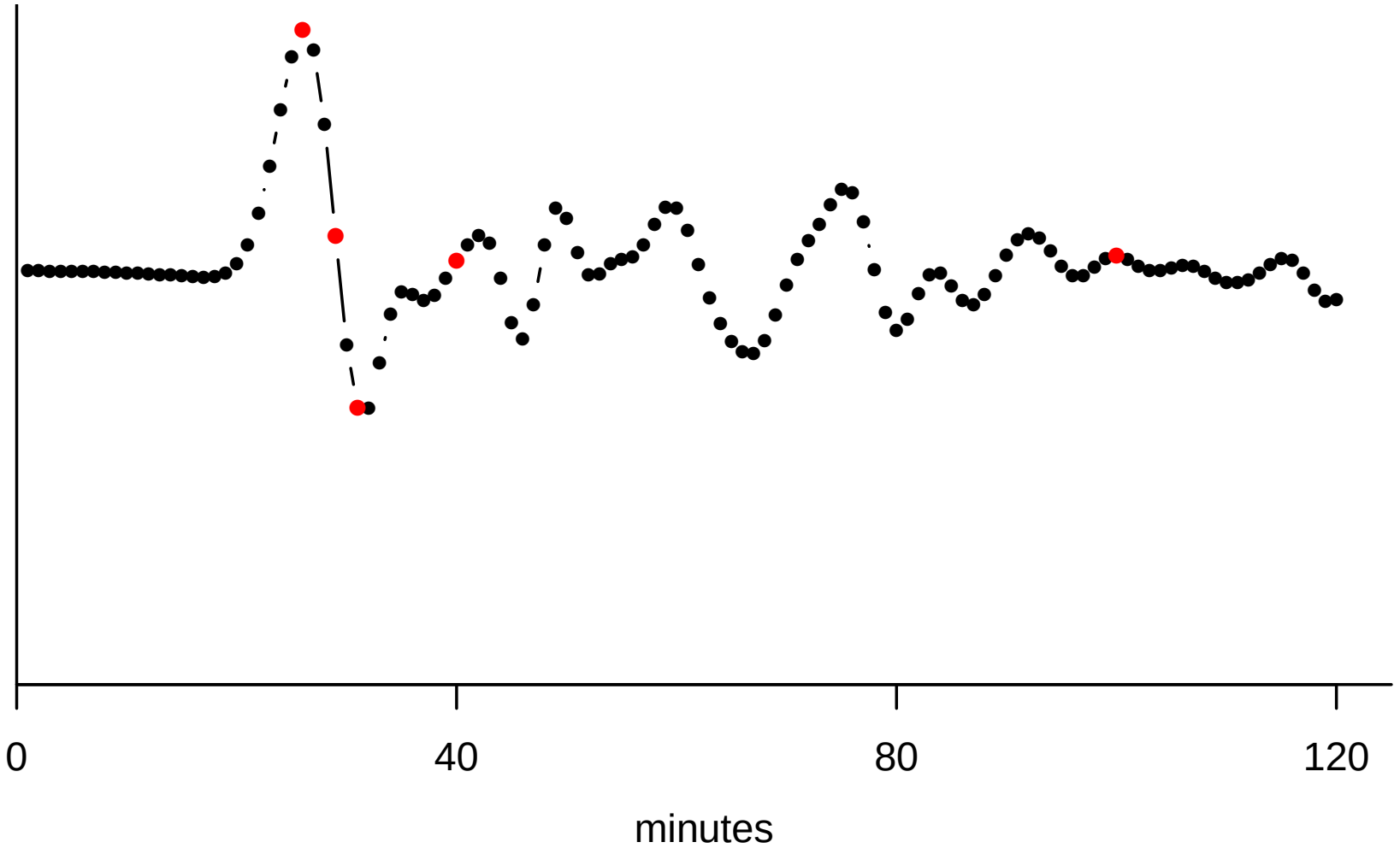
# Locations of 11 DART® Buoys and 42 Unit Sources



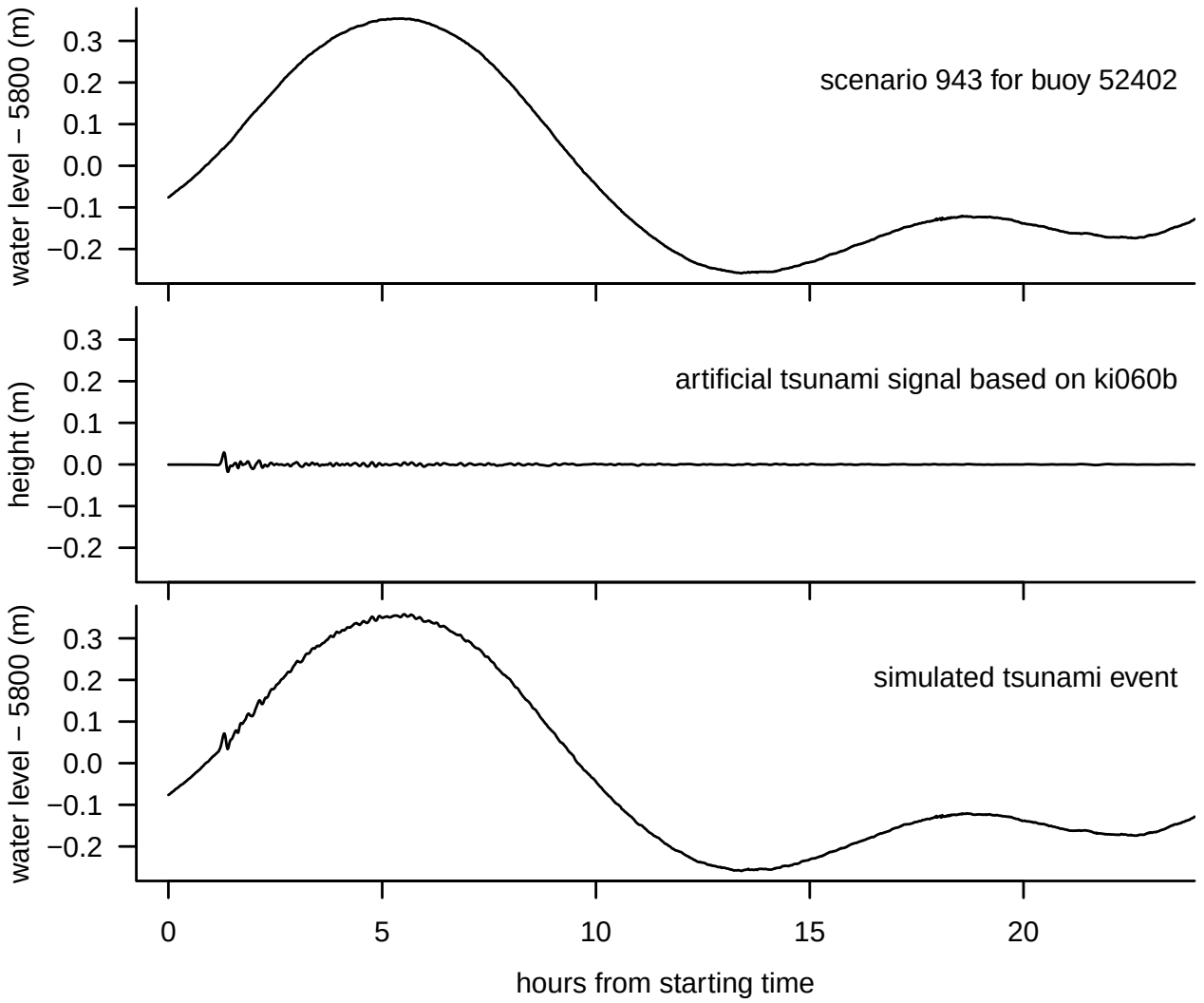
# Assessing Performance of Five Methods: IV

- next plot shows artificial tsunami signal for buoy ki060b paired with unit source ki060b
- five red dots mark arrival times of
  - first quarter wave
  - half wave
  - three-quarters wave
  - first full wave
  - one hour beyond first full wave

# Buoy 52402/Unit Source ki060b Pairing



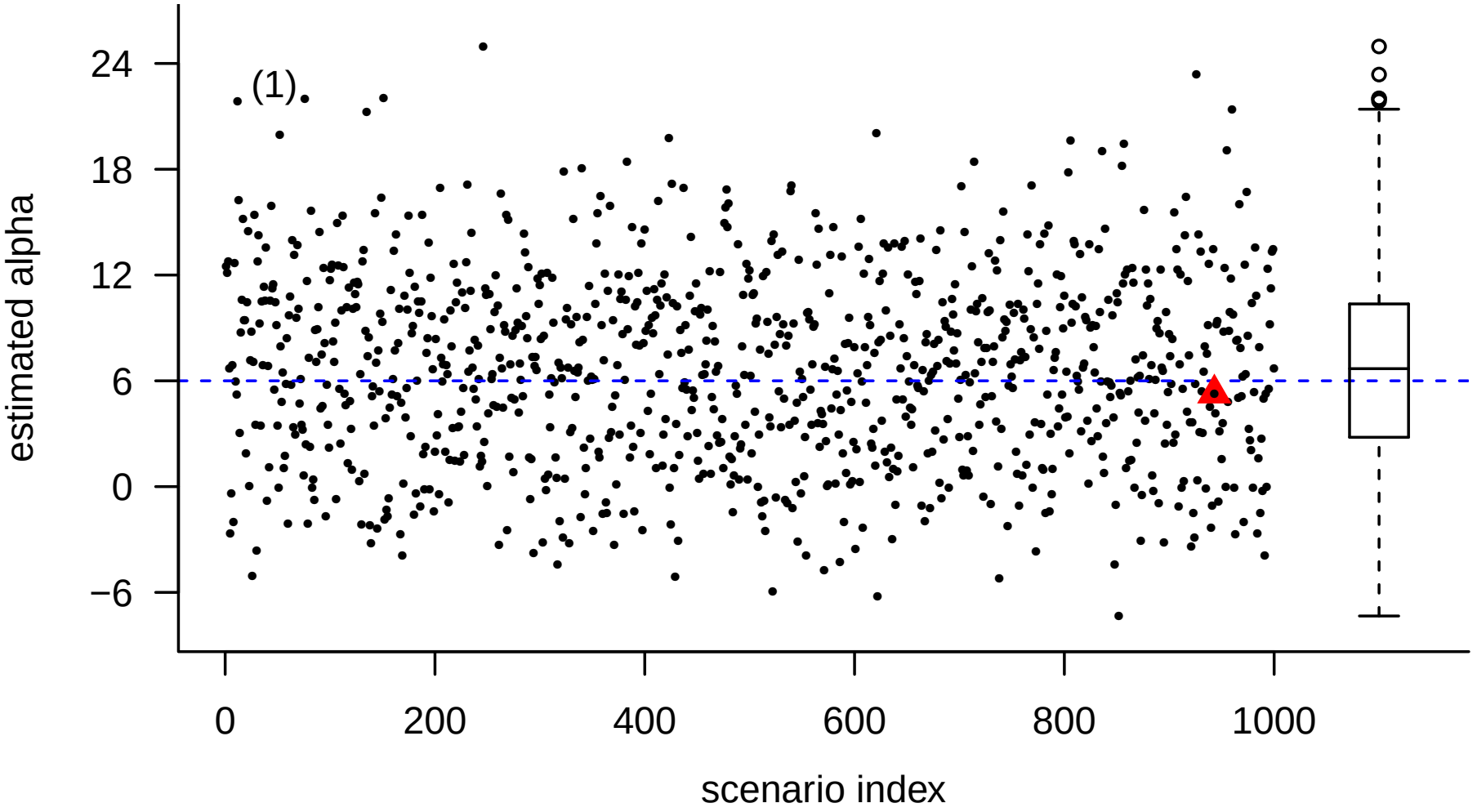
# Construction of Simulated Tsunami Event



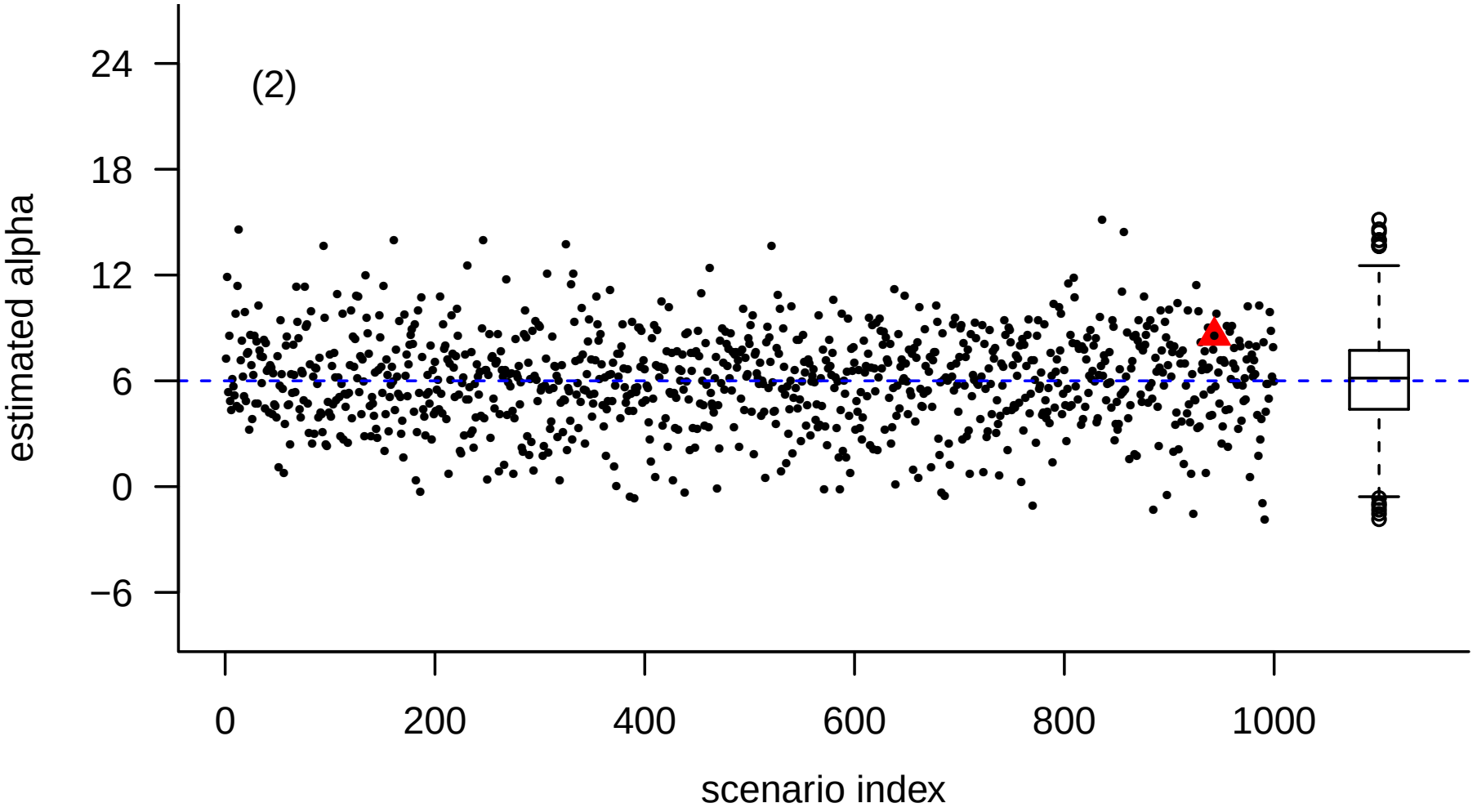
# Assessing Performance of Five Methods: V

- next set of plots show $\hat{\alpha}$'s estimated using OLS for five methods using
  - data up to 3/4 of first full wave
  - all 1000 scenarios for buoy 52402
  - artificial tsunami based on unit source ki060b
- triangles mark scenario 943

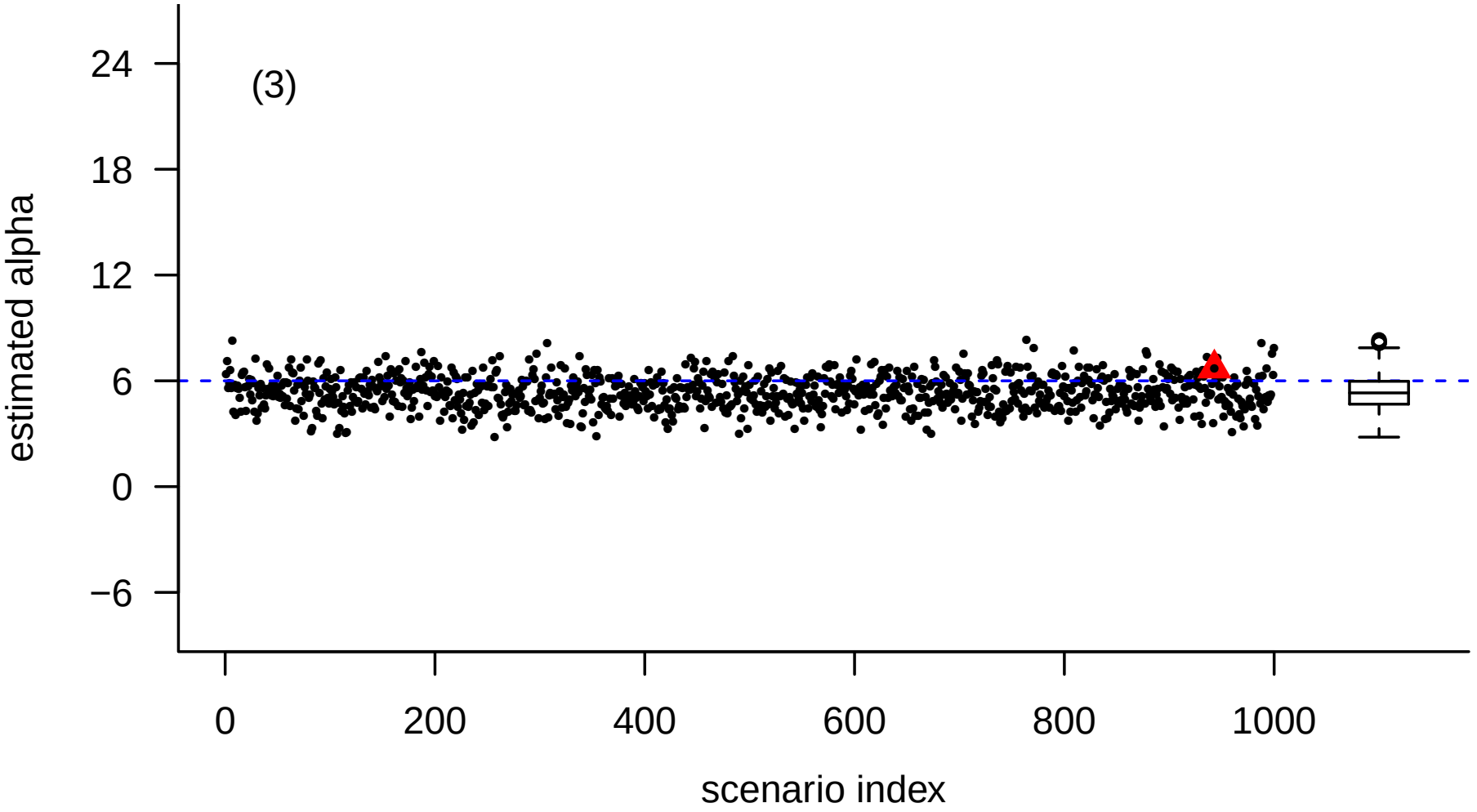
# $\hat{\alpha}$'s Using 29 Day Harmonic Analysis



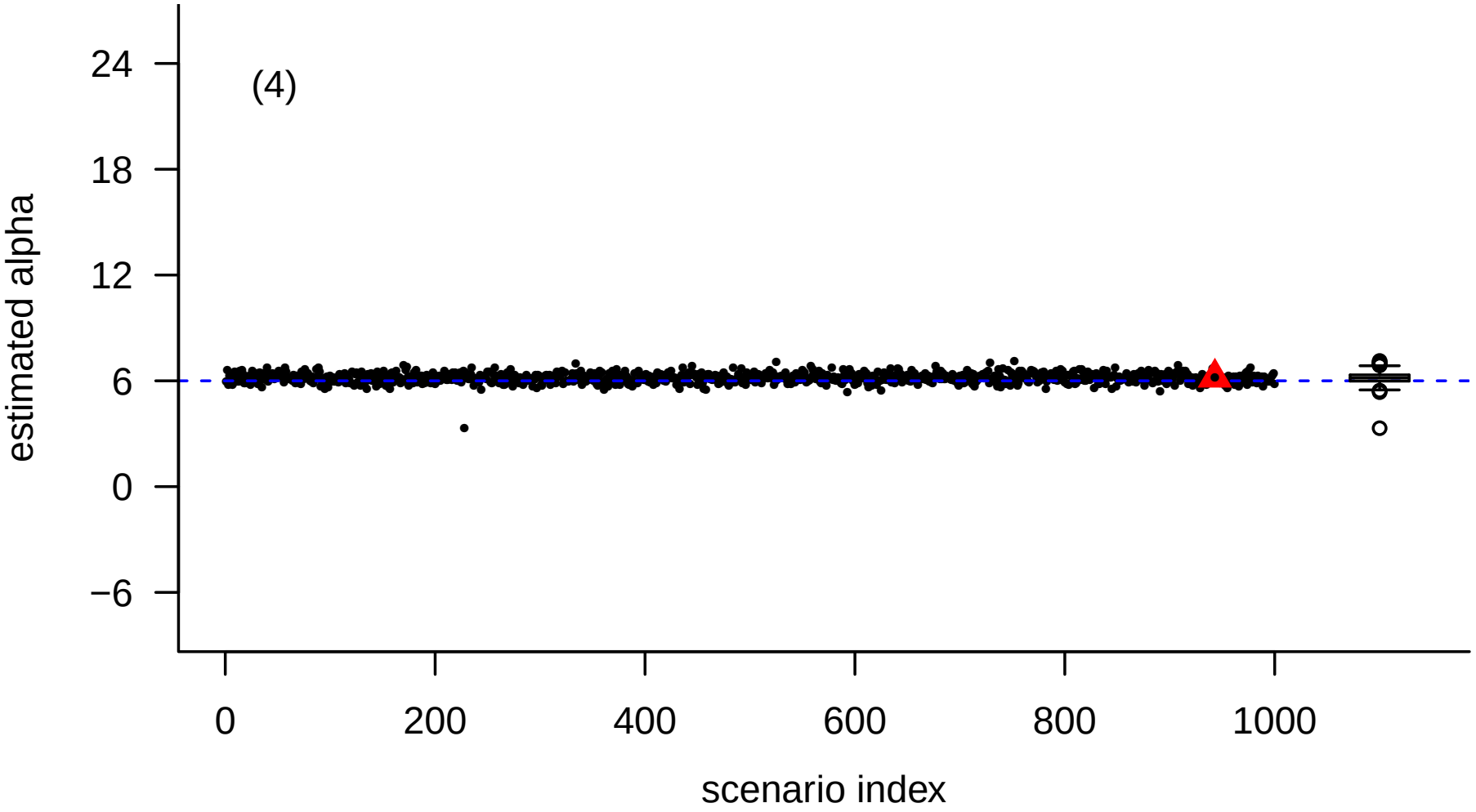
# $\hat{\alpha}$'s Using Long Harmonic Analysis



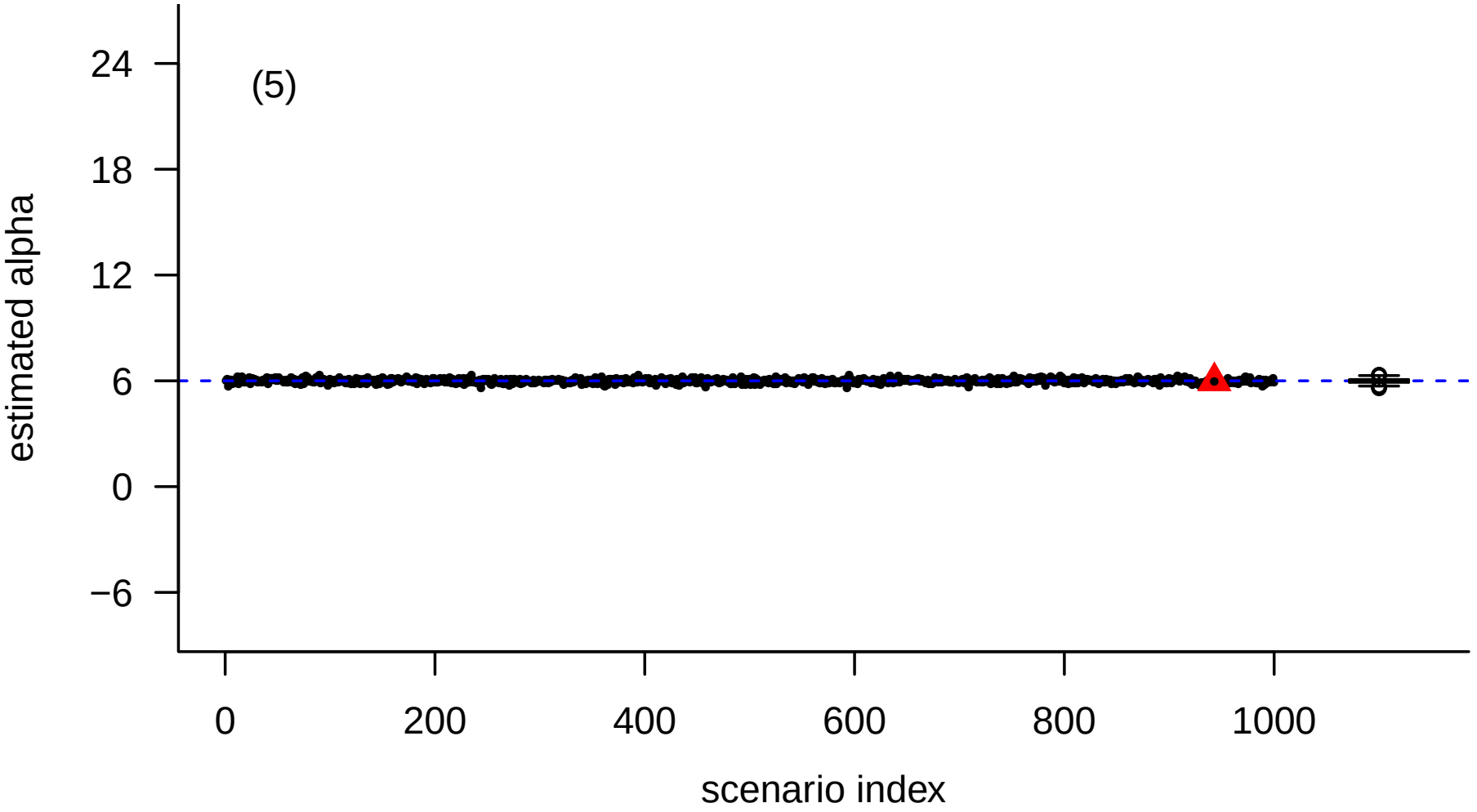
# $\hat{\alpha}$'s Using Empirical Orthogonal Functions



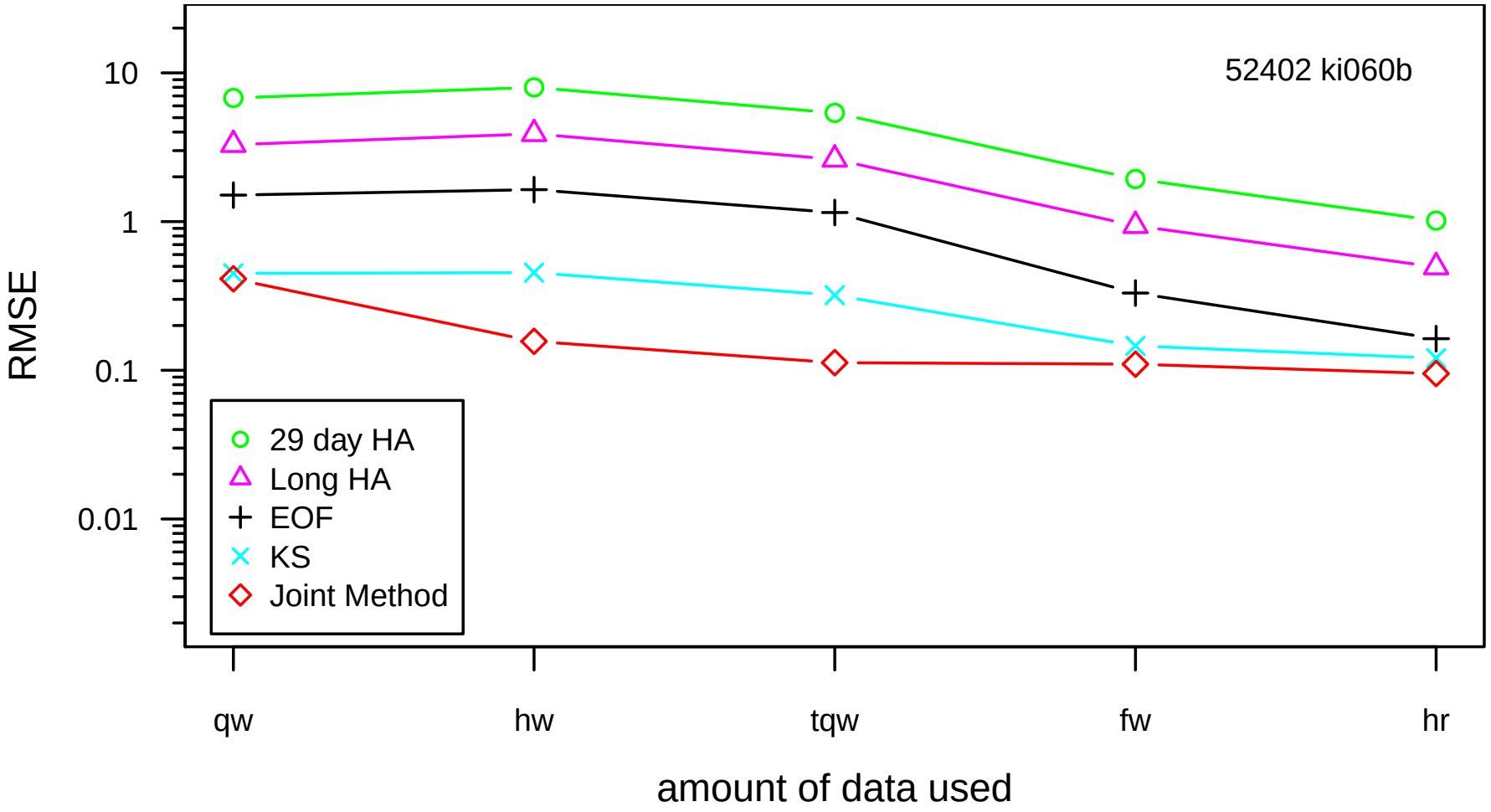
# $\hat{\alpha}$'s Using Kalman Smoothing



# $\hat{\alpha}$'s Using Joint Method



# Root-Mean-Square Errors for 1000 $\hat{\alpha}$ Estimates



## Assessing Performance of Five Methods: VI

- why does joint method work well?
  - really only need a simple model for tidal fluctuations when dealing with short stretches of time
  - least squares theory suggests that simultaneous estimation of model parameters is preferable to stage-wise estimation if predictors are correlated
- why might joint method fail?
  - method needs an appropriate model for tsunami signal, but sweeping elastic net offers a promising solution to this potential drawback
  - long stretches of time problematic since simple model for tidal fluctuations can deteriorate ($M = 1$ versus $M = 2$ needs more study)

# Concluding Comments: I

- sweeping elastic net soon to be implemented within SIFT to help operators at tsunami warning centers select appropriate model for tsunami signal during ongoing event (manual method currently in use too time consuming)
- many statistical issues remain, including following three

1. selection of blocksize for sweeping elastic net
   - used $4 \times 3$ in example here, but choice somewhat arbitrary
   - tsunami warning officers comfortable with setting size based on magnitude of earthquake (at least to begin with)

# Concluding Comments: II

2. windowing of buoy data to use for model fitting
   - joint method impacted because simple tidal model deteriorates as window width increases
   - geophysical models typically get first full wave of tsunami signal OK, but do not do well after that
3. assessing uncertainty in coastal inundation forecasts
   - how does uncertainty in estimating $\boldsymbol{\alpha}$ translate into uncertainty in forecasts?

# References: I


- A.E. Hoerl and R.W. Kennard (1970), 'Ridge regression: biased estimation for nonorthogonal problems', *Technometrics* vol. 12, no. 1, pp. 55–67
- Y. Okada (1985), 'Surface deformation due to shear and tensile faults in a half-space', *Bulletin of the Seismological Society of America*, vol. 75, no. 4, pp. 1135–1154
- D.B. Percival, D.W. Denbo, M.C. Eblé, E. Gica, P.Y. Huang, H.O. Mofjeld, M.C. Spillane, V.V. Titov and E.I. Tolkova (2015), 'Detiding DART® buoy data for real-time extraction of source coefficients for operational tsunami forecasting', *Pure and Applied Geophysics*, vol. 172, no. 6, pp. 1653–1678
- D.M. Percival, D.B. Percival, D.W. Denbo, E. Gica, P.Y. Huang, H.O. Mofjeld and M.C. Spillane (2014), 'Automated tsunami source modeling using the sweeping window positive elastic net', *Journal of the American Statistical Association*, vol. 109, no. 506, pp. 491–499
- R. Tibshirani (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288
- E. Tolkova (2010), 'EOF analysis of a time series with application to tsunami detection', *Dynamics of Atmospheres and Oceans*, vol. 50, pp. 35–54

# References: II


- H. Zou and T. Hastie (2005), 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society, Series B*, vol. 67, pp. 301–320