A Bayesian Approach to Estimate Weights for GMPE Models in a Logic Tree

Natural Phenomena Hazards

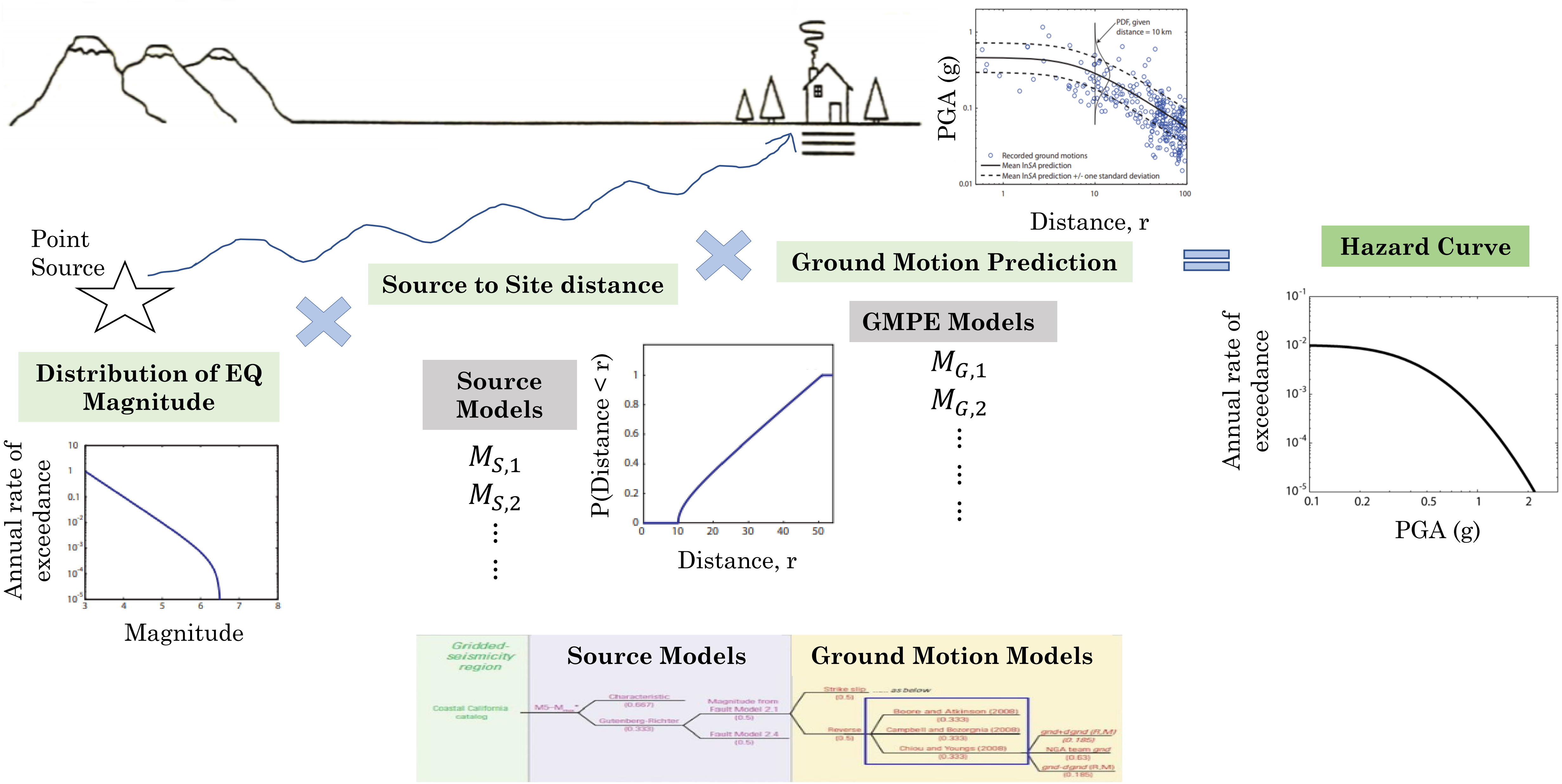
Saran Bodda **(NCSU), Abhinav Gupta (NCSU), Merlin Keller (EDF), Gloria Senfaute (EDF)**

Électricité de France (EDF)

Center for Nuclear Energy Facilities and Structures (CNEFS)
North Carolina State University
Raleigh, NC 27695-7908

CNEFS

Probabilistic Seismic Hazard Analysis (PSHA)



Model Uncertainty

- ***"Essentially all models are wrong, but some are useful."*** – *George E. P. Box*
- It depends on how accurately a mathematical model describes the true system for a real-life situation, since models are almost always only approximations to reality.
- Model Selection – process of selecting one candidate model from among a collection of candidate models
    - predictions obtained with a single model can lead to statistical bias or underestimation of the output variability
- Model Averaging (ensemble modeling)

Ensemble Modeling

- A process where multiple diverse models are created to predict an outcome, either by using many different modeling algorithms or using different training data sets.
- The ensemble model then aggregates the prediction of each candidate model and results in once final prediction for the unseen data.
- The motivation for using ensemble models is to reduce the generalization error of the prediction.
- One of the most common challenge with ensemble modeling is to find optimal weights to ensemble candidate models.
- For some applications, we assume equal weight (or expert judgement) for all models and take the average of predictions.

Bayesian Model Averaging (BMA)

- BMA approach refers to the process of estimating a desired quantity of interest (y) under each candidate model (M_k) and then averaging the estimates based on how likely each model is, in a list of candidate models.

$$p(y \mid D) = \sum_{k=1}^{K} p(y \mid M_k, D)\, p(M_k \mid D)$$

- Posterior Probability (or weights) of each model

Model prior probability: ***uniform***

$$p(M_k \mid D) = w_k = \frac{p(D \mid M_k)\, p(M_k)}{\sum_{k=1}^{K} p(D \mid M_k)\, p(M_k)}$$

- Marginal Likelihood of candidate model

$$p(D \mid M_k) = \int_{\theta_k} p(D \mid \theta_k, M_k)\, p(\theta_k \mid M_k)\, d\theta_k$$

Statistical Model 1

General model

$$y_i = f_k(x_i,\beta) + \varepsilon_{k,i}$$

Stochastic error

$$\varepsilon_{k,i} \sim N\left(0,\sigma_k^2\right)$$

y_i is the QoI of the considered database

f_k is the k-th candidate model

$f_k(x_i,\beta)$ is the QoI predicted by the k-th candidate model, given a vector x_i of regressors

$$p(y \mid D) = \sum_{k=1}^{K} p(y \mid M_k, D)\, p(M_k \mid D)$$

Fixed parameter setting

$$y_i = f_k(x_i,\beta) + \varepsilon_{k,i}$$

$$\sum_{i=1}^{N} \log\left(\sum_{k=1}^{K} w_k f_k(y_i \mid x_i,\beta)\right)$$

$f_k(x_i,\beta)$ is fixed. Weights can be estimated using Maximum likelihood approach

Statistical Model 2

Fixed parameter setting with added bias term

$$y_i = f_k(x_i, \beta) + \underbrace{\mu_k}_{Bias\ Term} + \varepsilon_{k,i} \qquad \varepsilon_{k,i} \sim N\left(0, \sigma_k^2\right)$$

Since $f_k(x_i, \beta)$ is fixed, the unknown parameters are $\theta_k = (\mu_k, \sigma_k^2)$

- Parameters and weights can be estimated using
  - Bayesian Linear models (conjugate prior approach) without any approximations
  - MCMC methods with approximations

Statistical Model 3

Uncertain parameter setting

$$y_i = f_k(x_i, \beta) + \varepsilon_{k,i} \qquad \varepsilon_{k,i} \sim N\left(0, \sigma_k^2\right)$$

The unknown parameters are $\theta_k = (\beta_k, \sigma_k^2)$

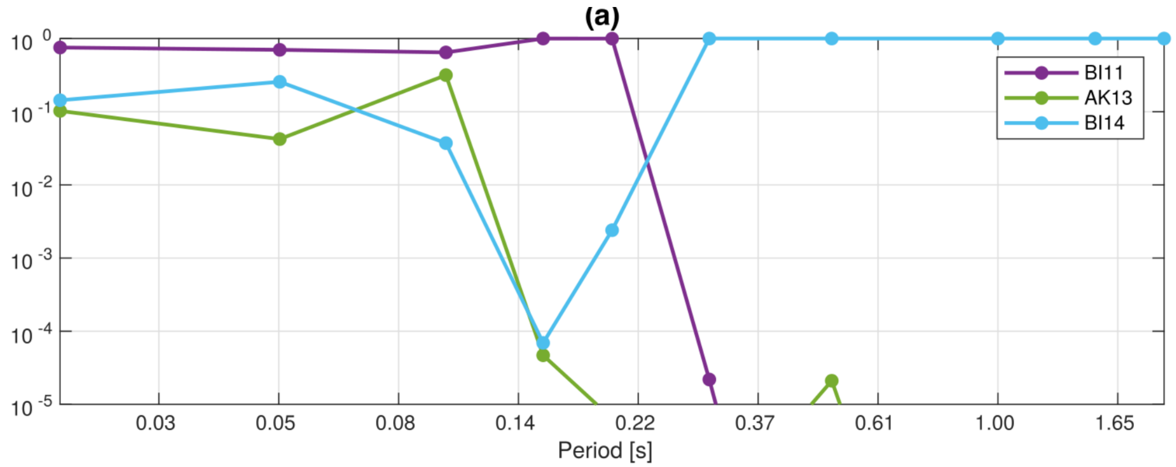
- Parameters and weights can be estimated using
  - Bayesian **Linear** models (conjugate prior approach) without any approximations
  - MCMC methods with approximations: adaptive Metropolis Hastings, sampling importance resampling with Laplace approximation

Bertin et al. (2019)

Using Bayesian model averaging to improve ground motion predictions

- Combines 9 GMPE models using logic-tree
- RESORCE-2013 database
- Update weights using Bayesian model averaging (BMA) approach
- BMA is implemented using two techniques: Markov chain Monte Carlo (MCMC) method and Maximum Likelihood estimation

$$y_i = f_k(x_i, \beta) + \mu_k + \varepsilon_{k,i}$$



Current Study

- 3 GMPE models from Bertin et al. (2019)
- pan-European Engineering Strong Motion (ESM)-2018 database
- Openquake and GMPE-smtk toolkit
- BMA is implemented using conjugate prior approach – Bayesian Linear Models

- Bindi11: https://github.com/gem/oq-engine/blob/master/openquake/hazardlib/gsim/bindi_2011.py
- Akkar14: https://github.com/gem/oq-engine/blob/master/openquake/hazardlib/gsim/akkar_2014.py
- Bindi14: https://github.com/gem/oq-engine/blob/master/openquake/hazardlib/gsim/bindi_2014.py

pan-European Engineering Strong Motion (ESM) flatfile

ESM Comprehensive Reference Table

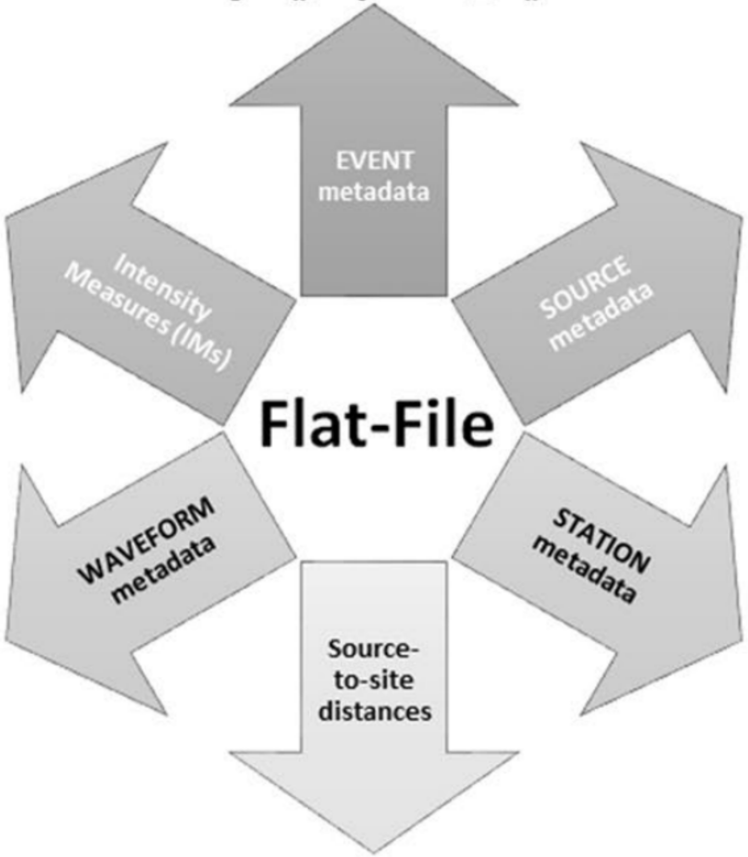
The flat file (2nd version, March 2018) consists of:

- 23,014 recordings from 2179 earthquakes and 2080 stations from Europe and Middle-East (1969 – 2018).
- Magnitudes range from 3.5 – 8.0 (includes shallow active crustal and subduction zones).
- Spectral amplitudes (5% damping, acceleration and displacement response) are provided for 36 periods, in the interval 0.01–10 s.
- Moment magnitude, focal depth, several distance metrics, style of faulting and parameters for site characterization.

- event_id, event_time, ISC_ev_id, USGS_ev_id, INGV_ev_id, EMSC_ev_id, ev_nation_code;
- ev_latitude, ev_longitude
- ev_depth_km
- fm_type_code
- M_L, M_w, M_s, EMEC_M_w

- X_pga, X_pgv, X_pgd
- X_T90
- X_Housner
- X_ia
- X_CAV
- X_Ty_yyy

- event_source_id
- es_strike, es_dip, es_rake
- es_z_top
- es_length, es_width



- instrument_type_code
- late_triggered_flag_01
- X_Channel_code
- X_Azimuth_deg
- X_hp, X_lp

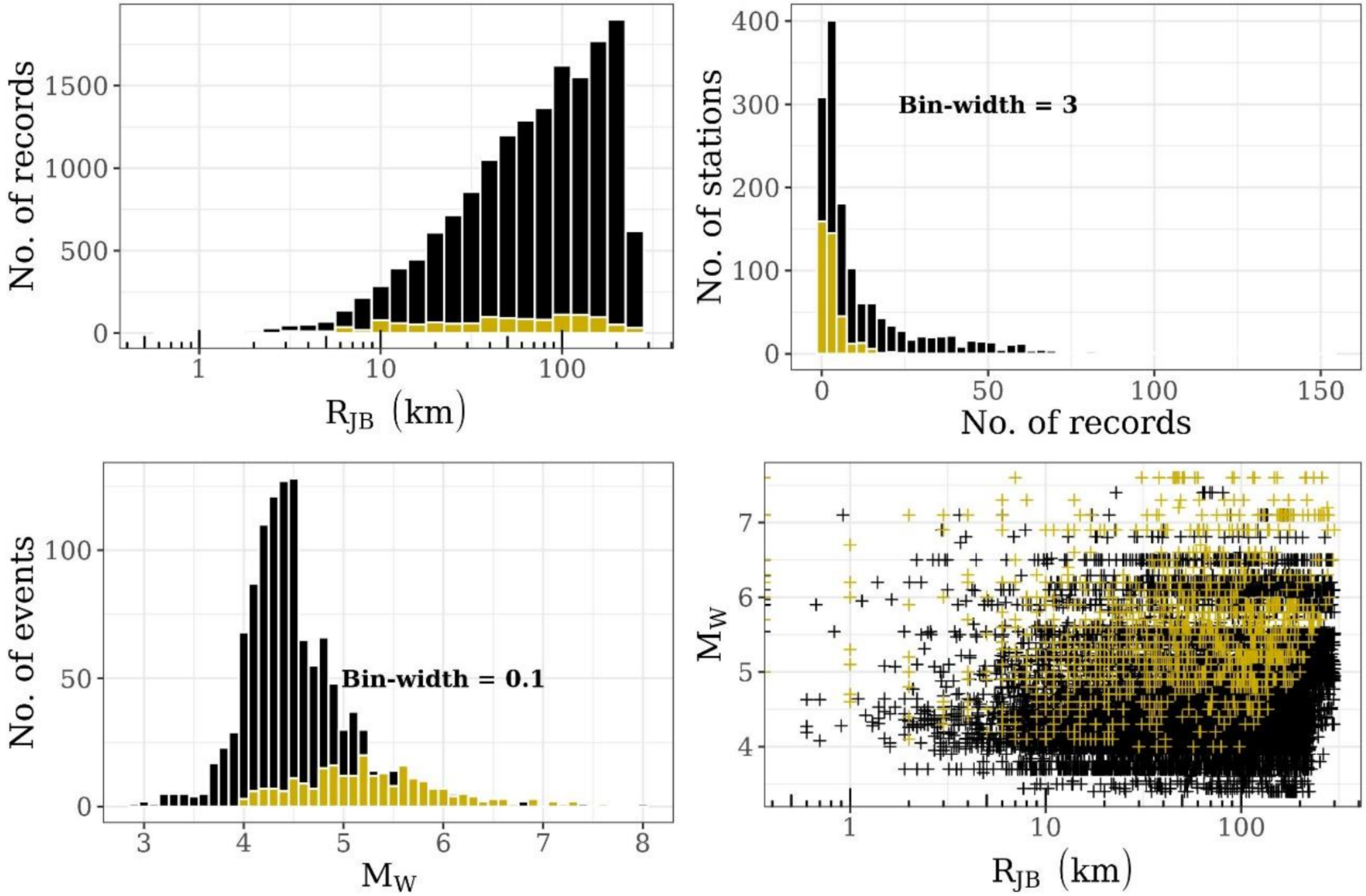
- network_code, station_code and location_code
- instrument_code
- sensor_depth_m
- Proximity_code, housing_code, installation_code
- st_nation_code
- st_latitude, st_longitude
- st_elevation
- ec8_code, ec8_code_method
- vs30_m_sec, vs30_calc_method, vs30_meas_type
- slope_deg, vs30_m_sec_WA

- epi_dist
- epi_az
- JB_dist
- rup_dist
- RX_dist, Ry0_dist

Comparison with RESORCE

- **RESORCE database** – 5882 recordings from 1814 earthquakes
- **ESM database** – 23,014 recordings from 2179 earthquakes
- Common records are about 2000, relative to 715 events.
- About 3000 records from RESORCE are not included in ESM
    - missing significant parts of the signal (e.g. S-wave arrivals)
    - missing one of the three components
    - stations without coordinates
- 46 events still not have an estimate of M_w in both databases

Comparison with RESORCE (SIGMA2 project)



RESORCE (yellow) and ESM (black)

Bertin et al. (2019) – Selection Criteria

- RESORCE Database (5882 records)
- **Selection Criteria: 939 records – training (739), testing (200)**
- Moment magnitude [5 – 7.3]
- Distance [4 – 150 km]
- V_{S30} velocity [300 – 1200 m/s]
- Fault mechanism – *Normal, Strike-Slip and Reverse*
- Geometric mean of horizontal spectral Acceleration is computed for 10 periods

	1	2	3	4	5	6	7	8	9	10
Period [s]	0.02	0.05	0.1	0.15	0.2	0.3	0.5	1	1.5	2

0, 0.04

ESM Database (19197 records)

Selection Criteria: 2356 records – training (1650), testing (706)

Simple Case Study

True model

$$y_i = f(x_i) + \varepsilon_i$$

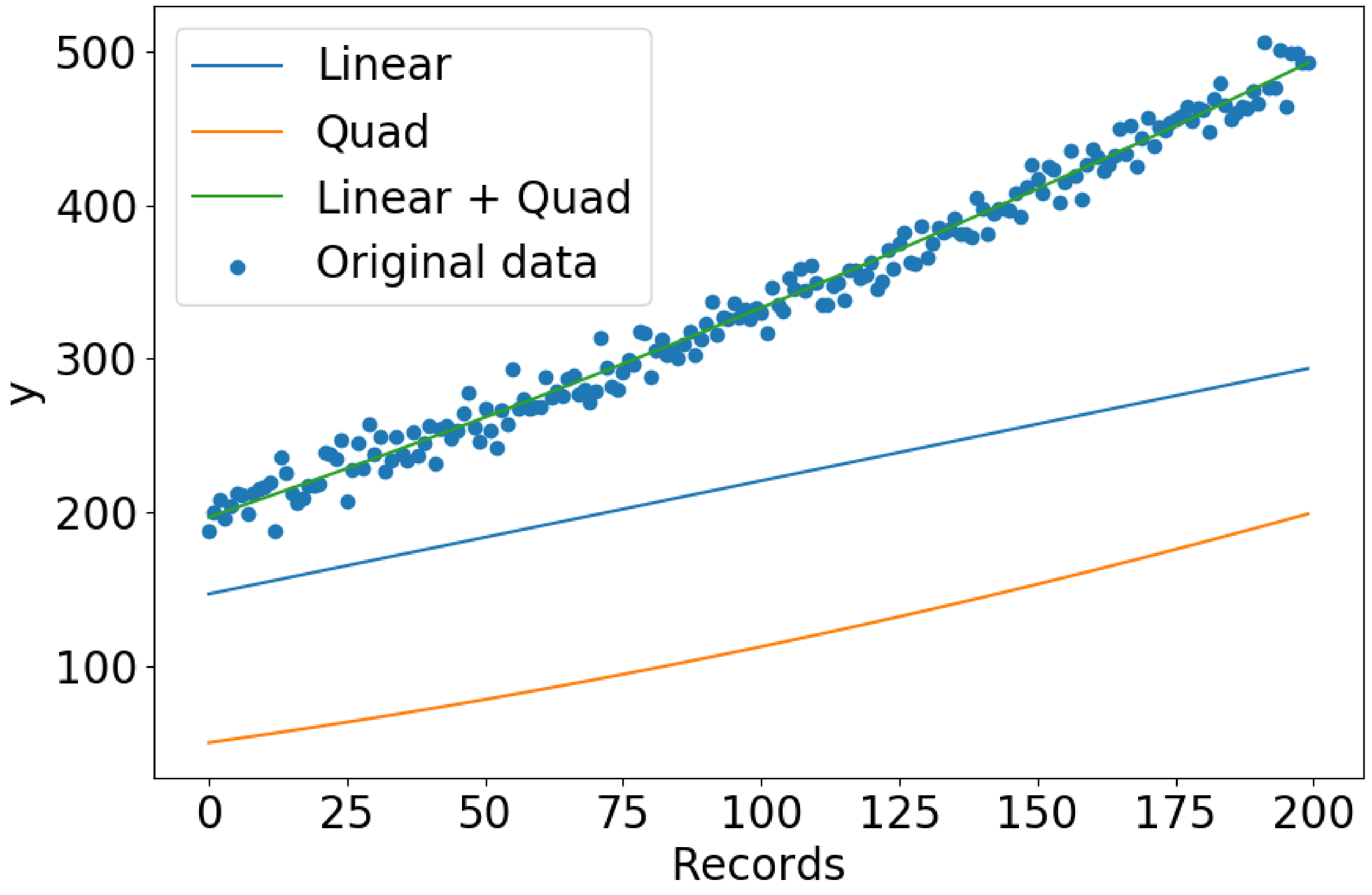
$$\varepsilon_i \sim N\left(0, 10^2\right)$$

$$f(x_i) = f_1(x_i) + f_2(x_i)$$

$$f_1(x_i) = -0.7 + 29.5x_i$$

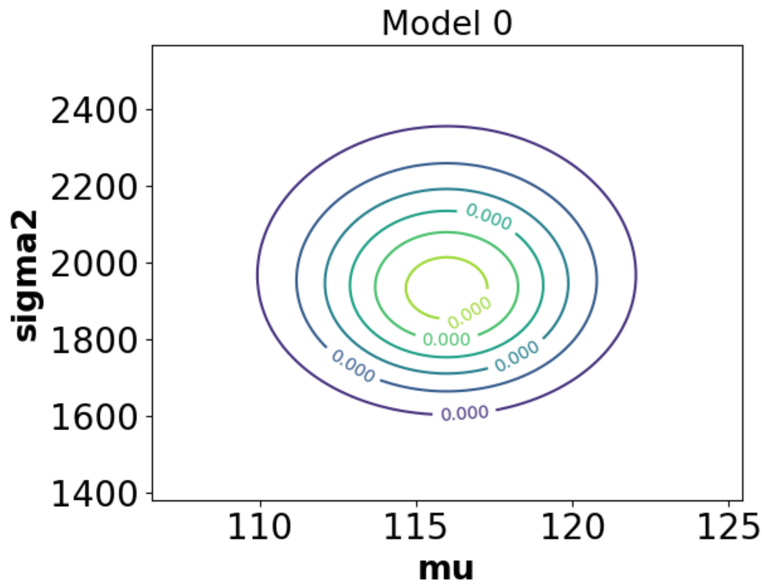
$$f_2(x_i) = 2x_i^2$$

$$f_3(x_i) = f_1(x_i) + f_2(x_i)$$

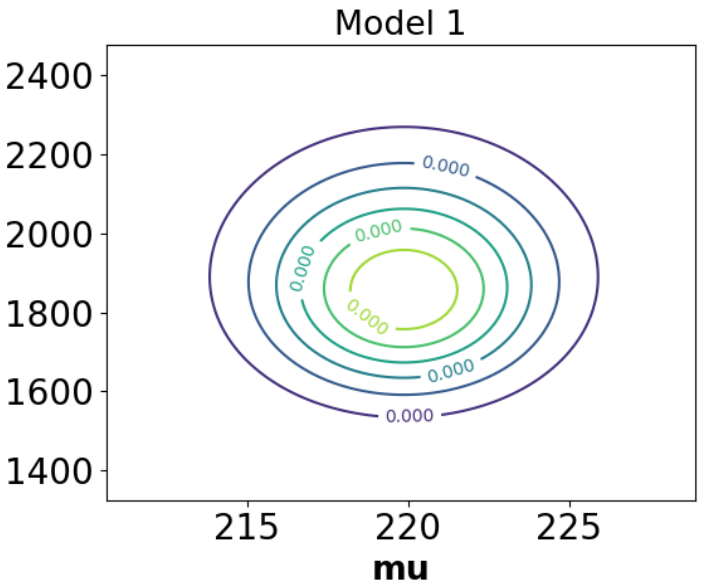


Posterior Marginal Distribution

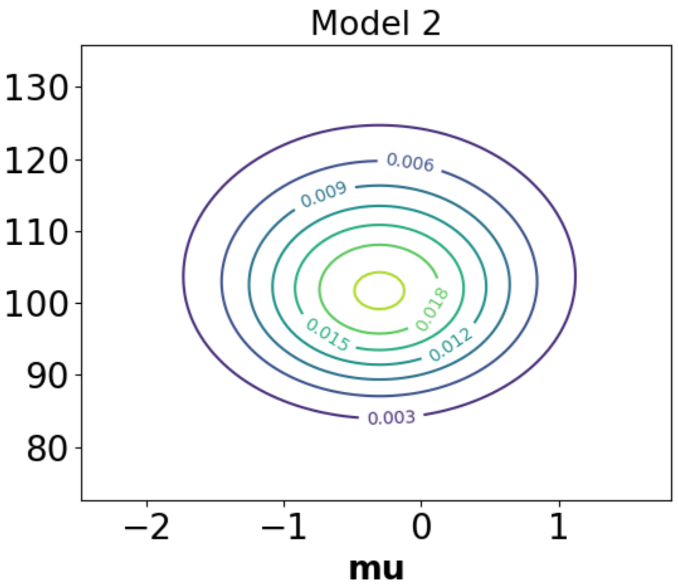
Linear



Quad



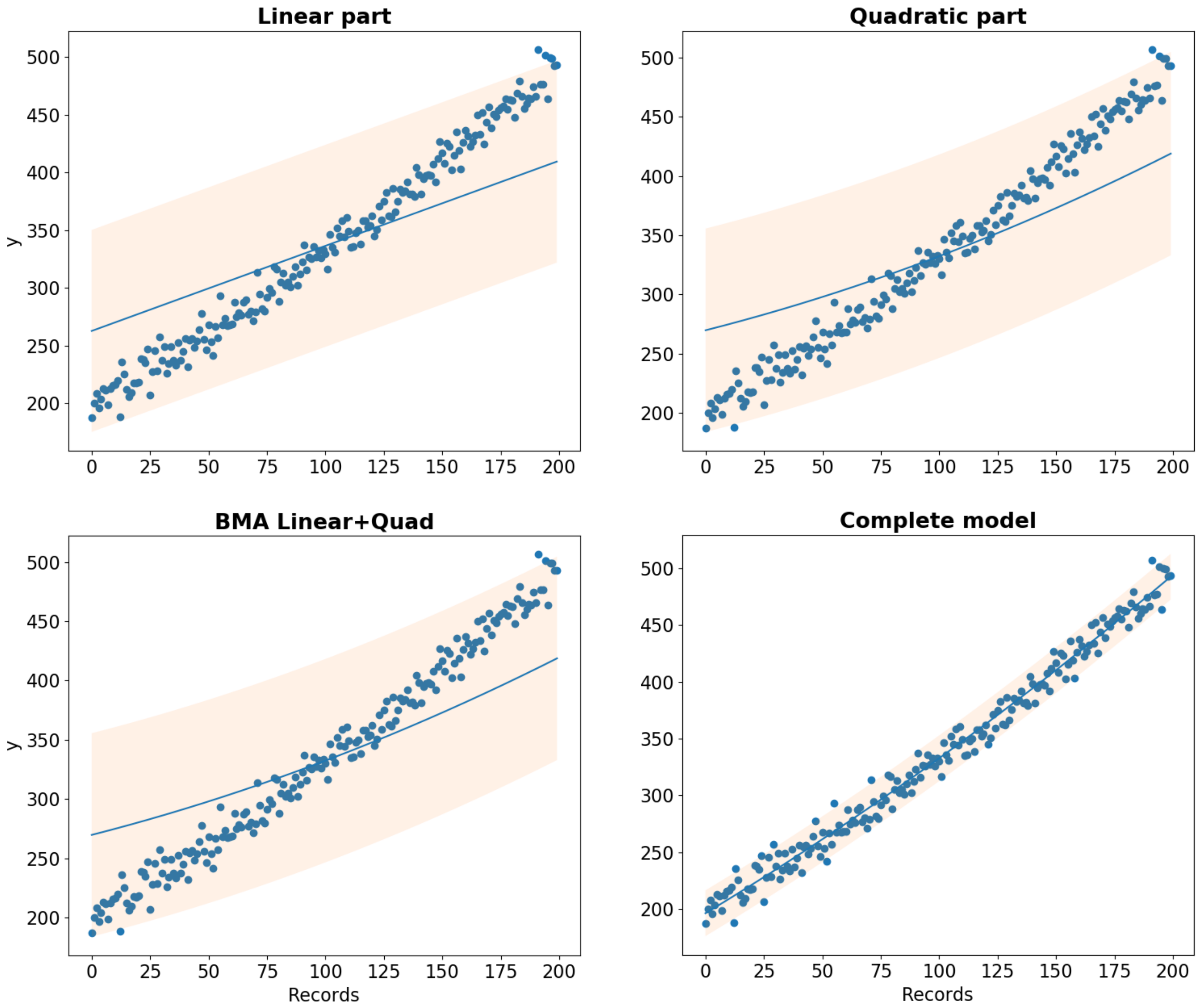
Linear + Quad



μ is concentrated around its true value (zero)

σ^2 is concentrated around its true value $\left(10^2\right)$

Bayesian Predictions



$$w_1 = 0.0175 \quad w_2 = 0.9825$$

$$w_1 = 1.41e-128$$

$$w_2 = 7.918e-127$$

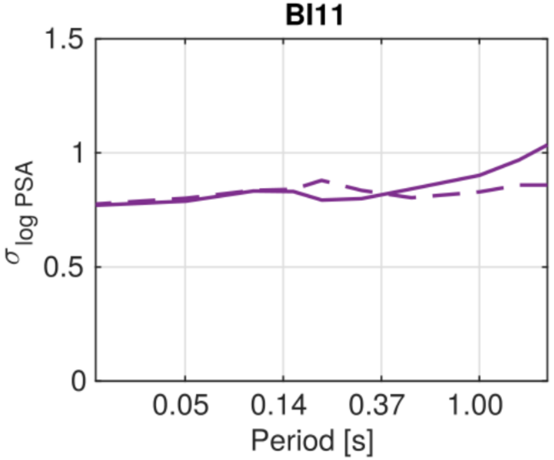
$$w_3 = 1$$

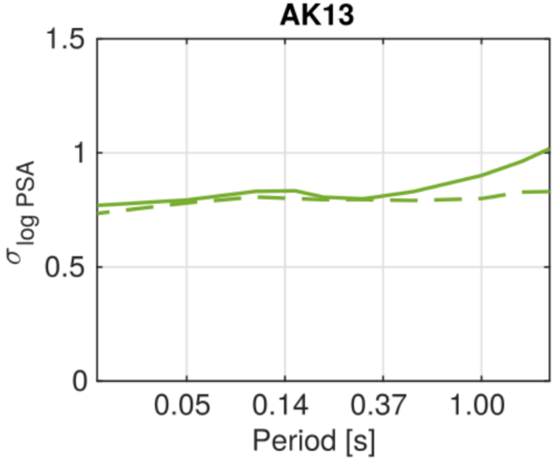
Comparison with Bertin et al. (2019)

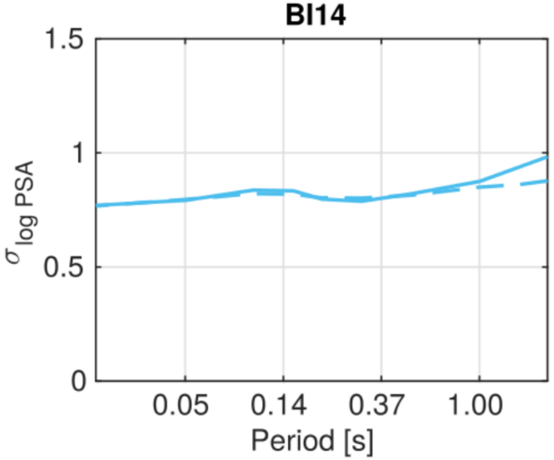
Root Mean Squared Error

$$\sigma_k^* = \left[\frac{1}{N}\sum_{i=1}^{N}\left(d_i - f_k(x_i) - \mu_k^*\right)^2\right]^{1/2}$$

RESORCE Database
MCMC



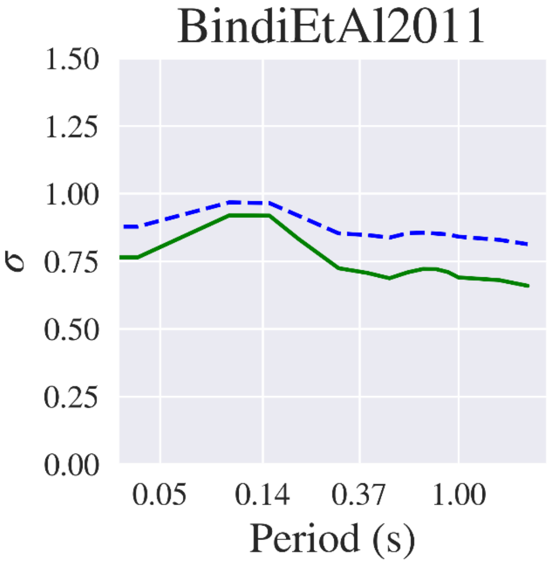


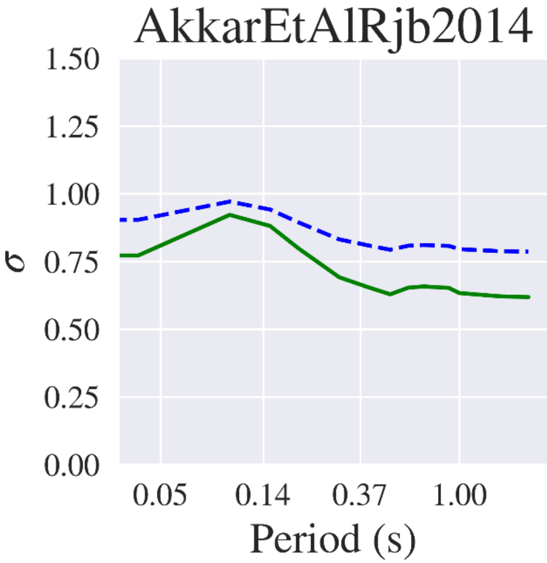


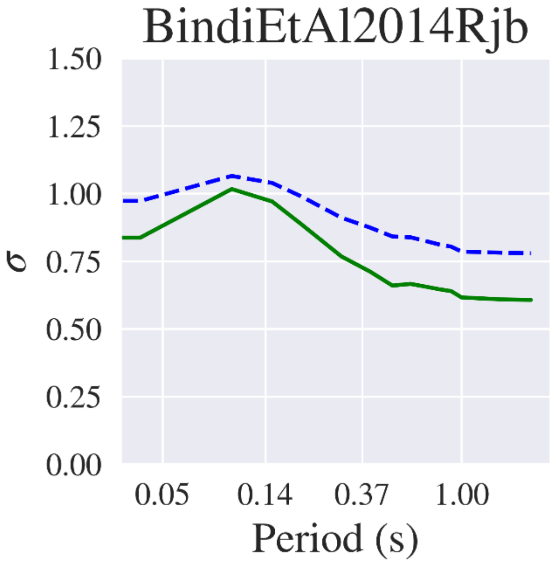
MLE method (solid lines)

Original values (dashed lines)

ESM Database
BLM







BLM method (solid lines)

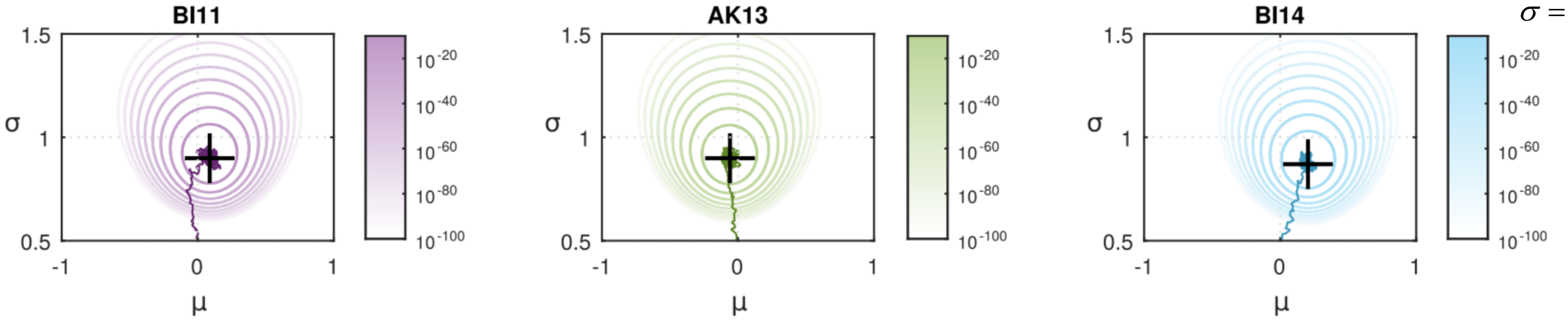
Original values (dashed lines)

Posterior Marginal Distribution

Priors

$\mu = U[-1,1]$

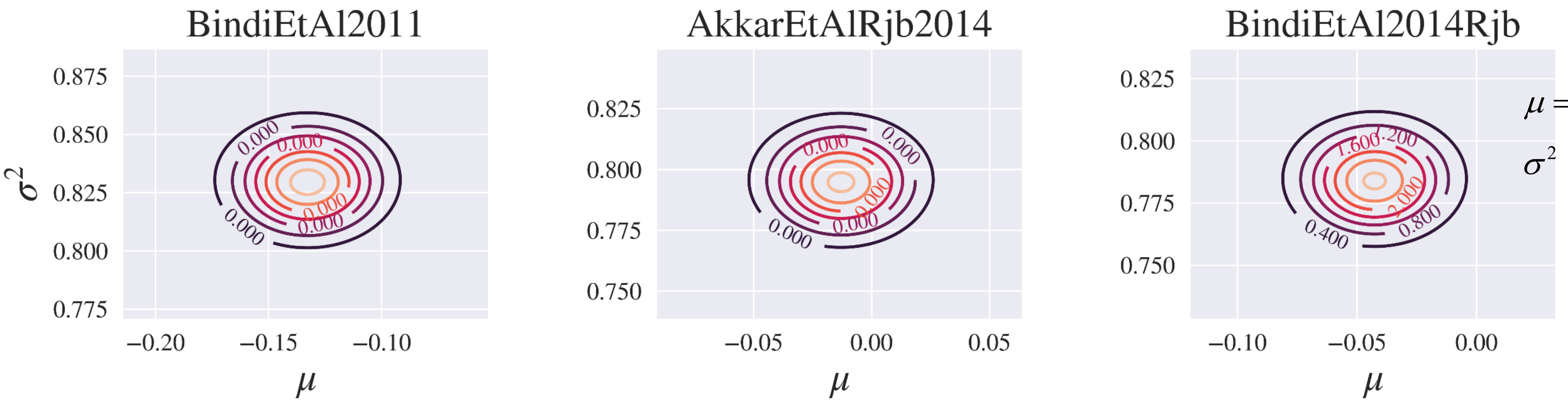
$\sigma = U[0.5,5]$



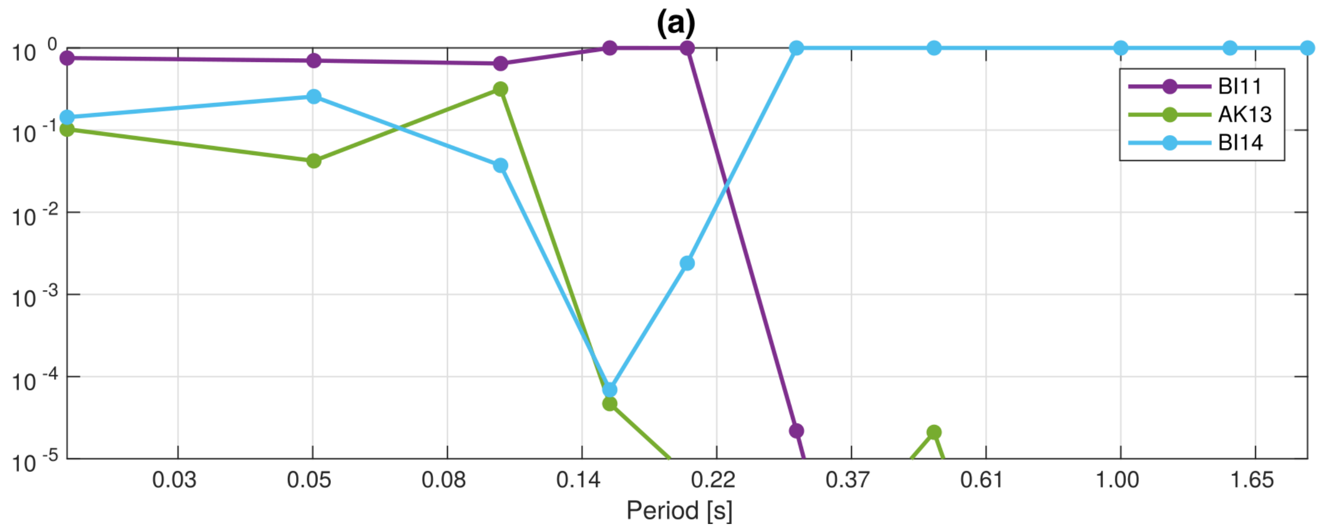
Priors

$\mu = N\left(0,10^2\right)$

$\sigma^2 = IG\left(0.001,0.001\right)$

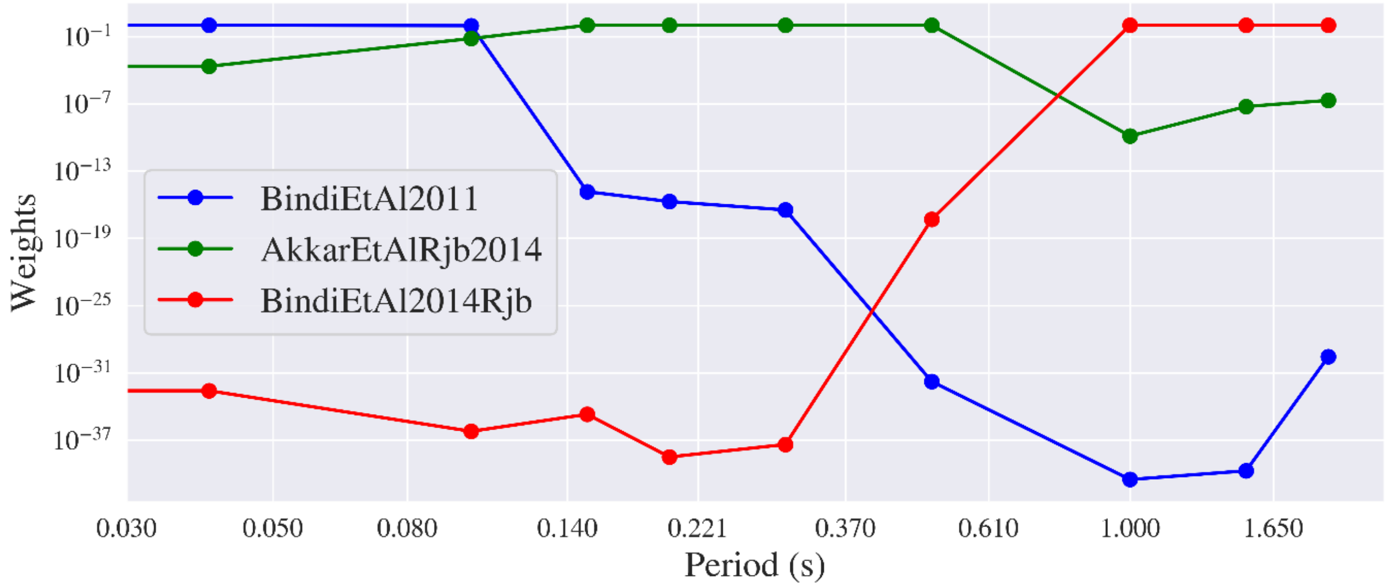


Model Weights



RESORCE Database

MCMC

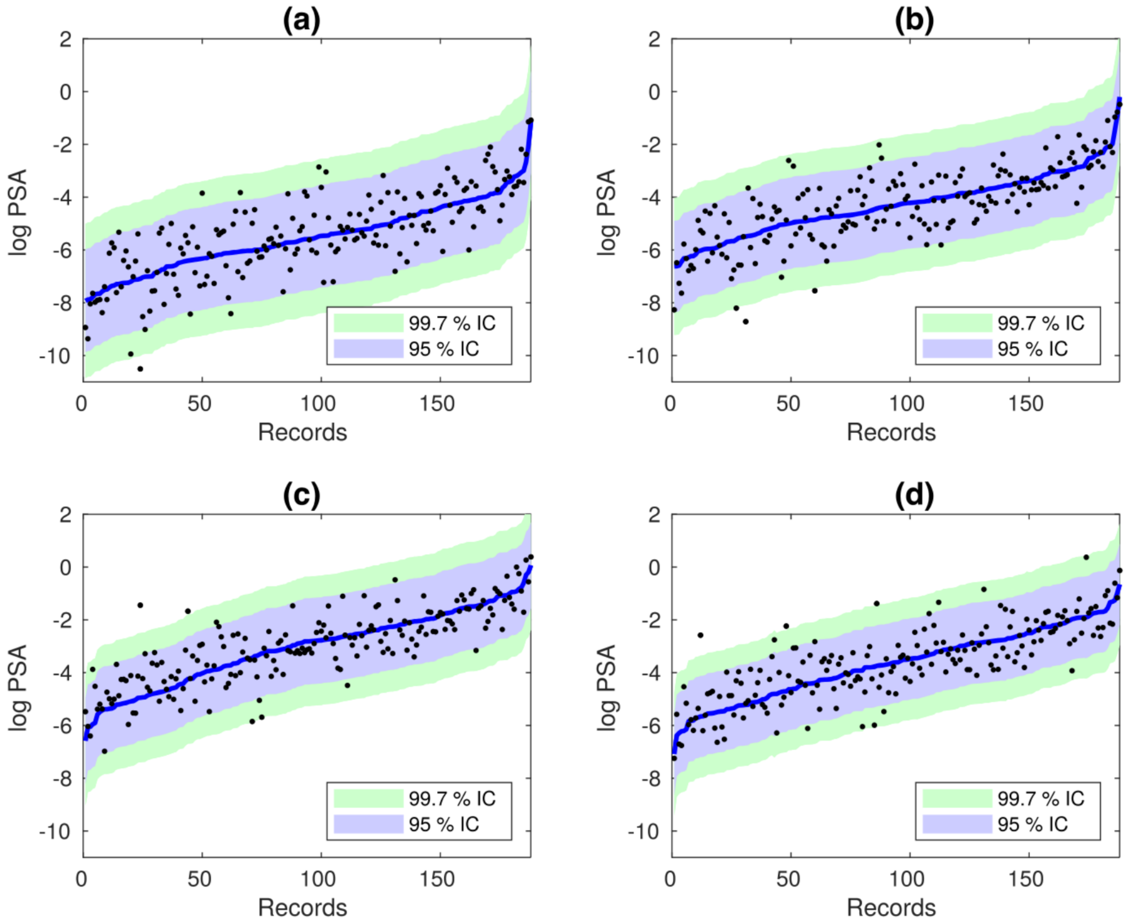


ESM Database

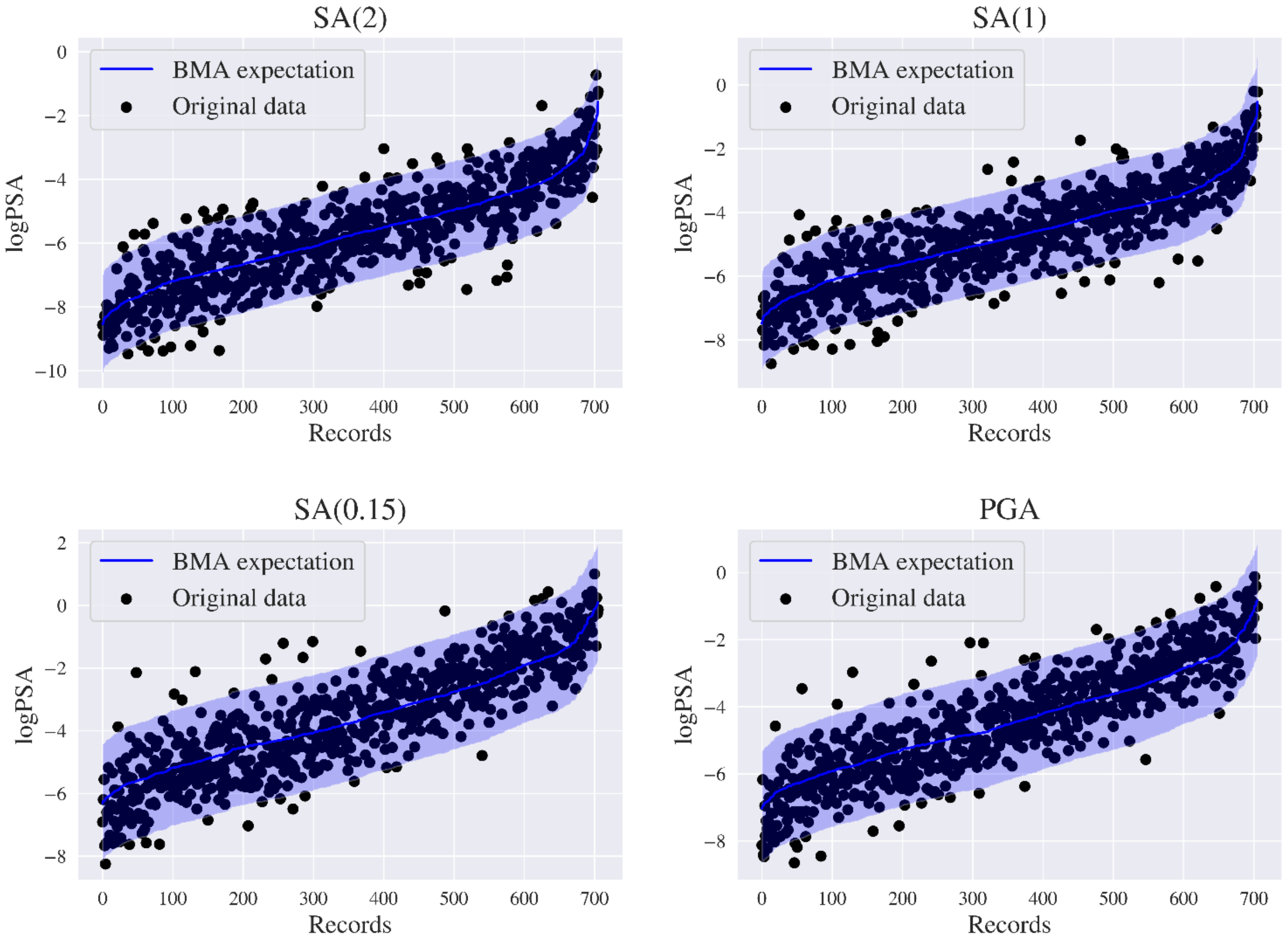
BLM

GMPE Predictions

Bertin et al. (2019) - RESORCE Database

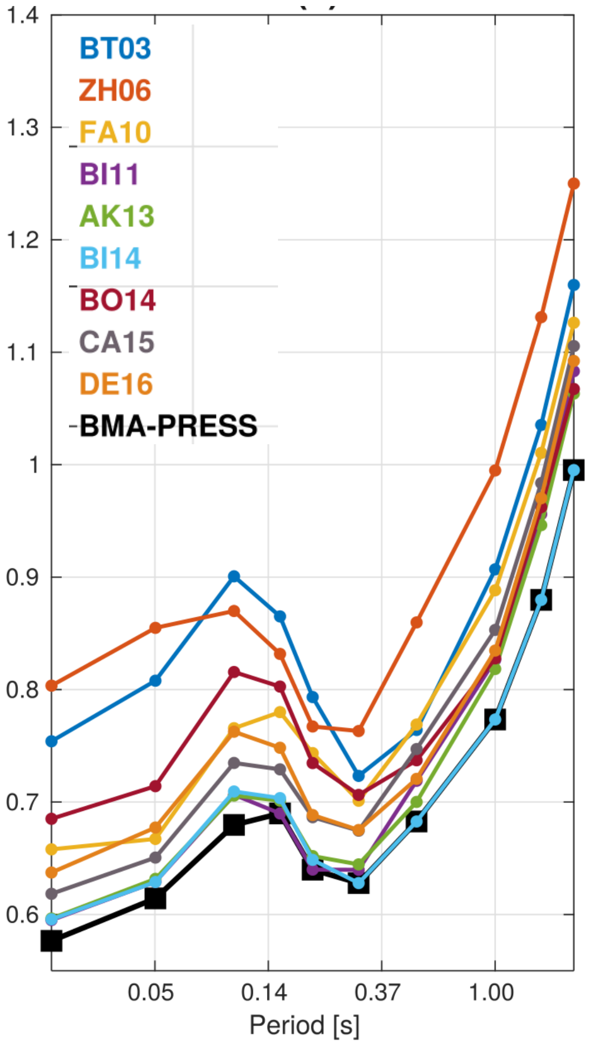


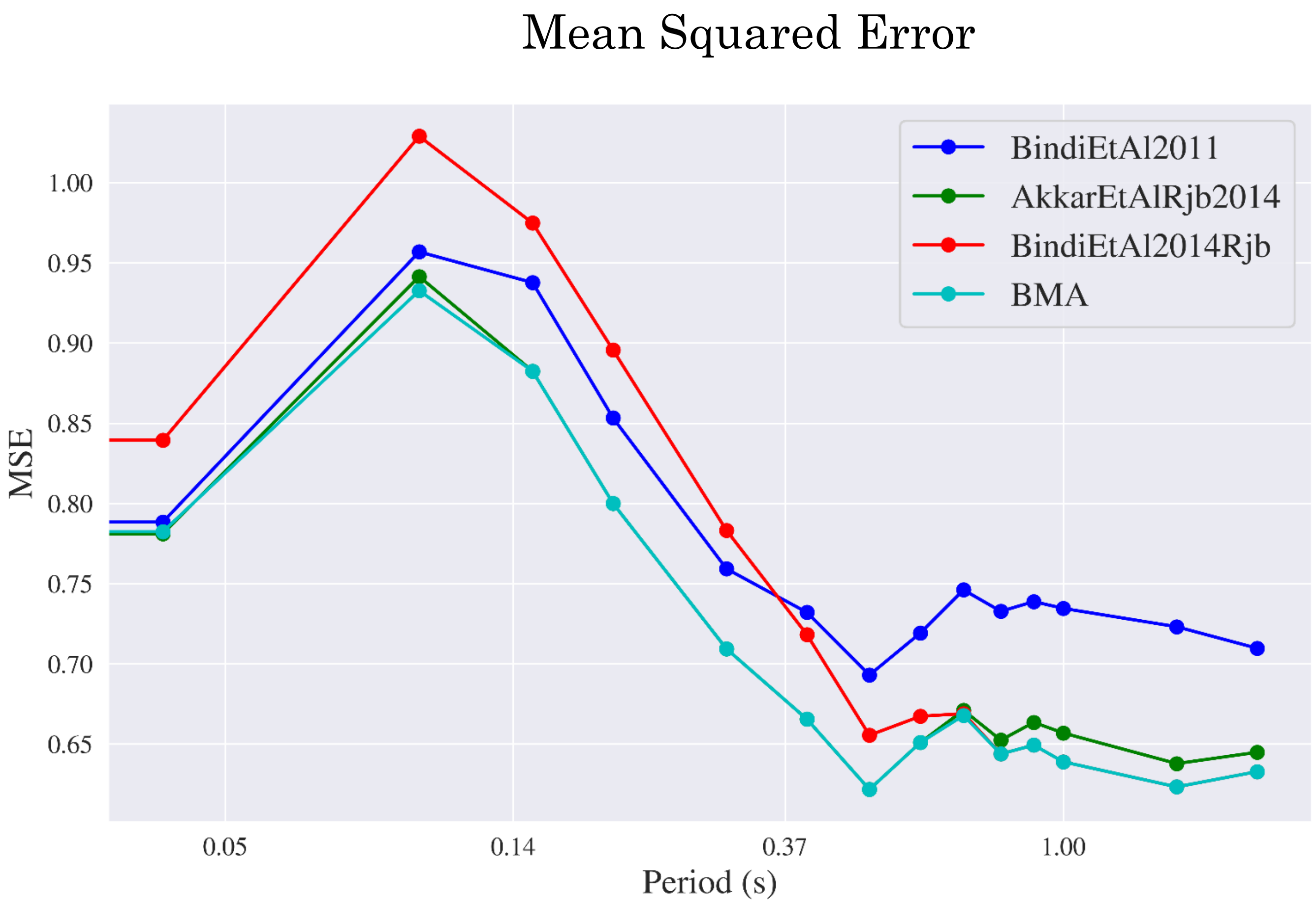
BLM - ESM Database



Testing

$$\text{PRESS}_{Bertin} = \frac{1}{N}\sum_{i=1}^{N}\left[E\left(y\left(x_n\right)\right)-d_n\right]^2$$





Conclusions

- Statistical methodology for updating weights without approximations.
- Calibrate bias along with standard deviation.
- Rank, select and combine models without expert judgement.
- Improved predictive performance for testing data set.