Statistical tests for evaluating predictability experiments in Japan

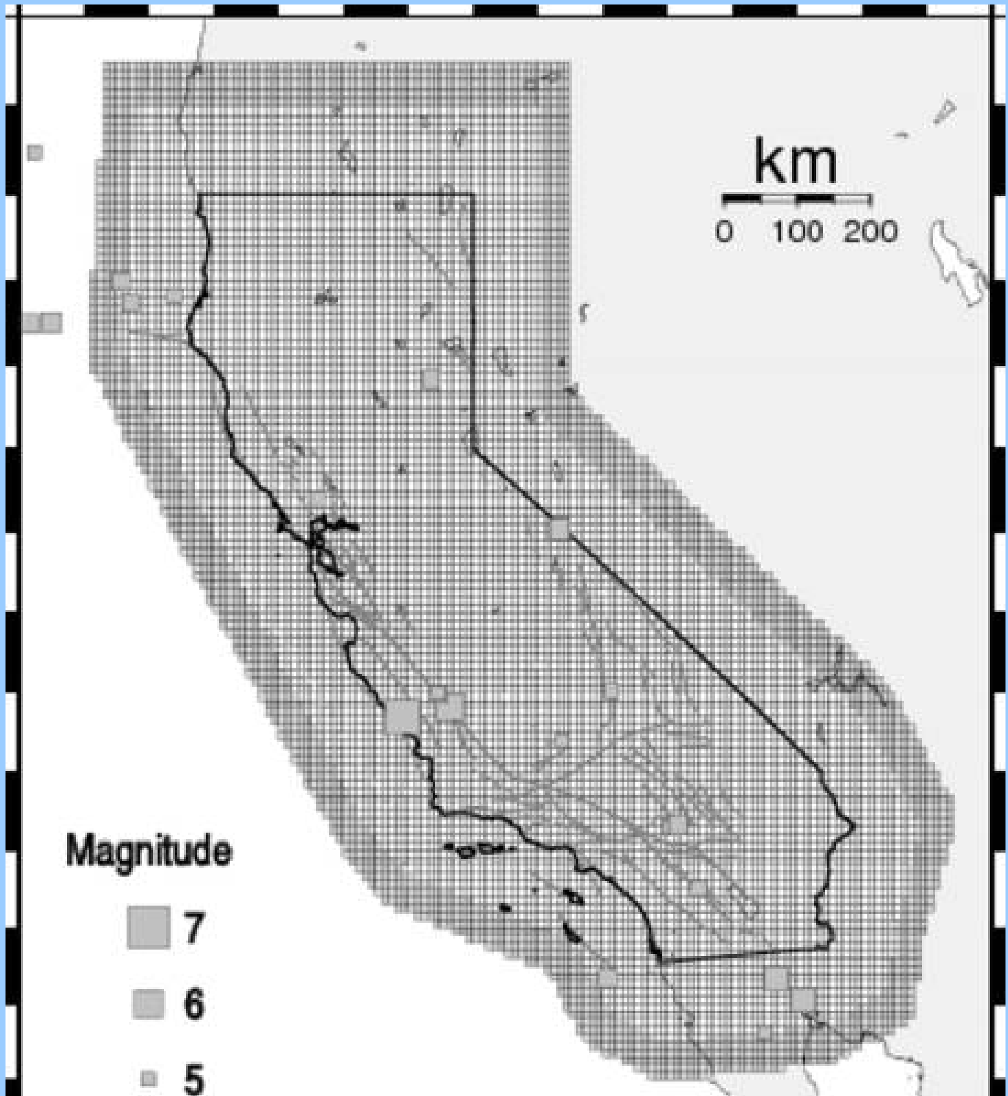
Jeremy Douglas Zechar
Lamont-Doherty Earth Observatory of Columbia University

Outline

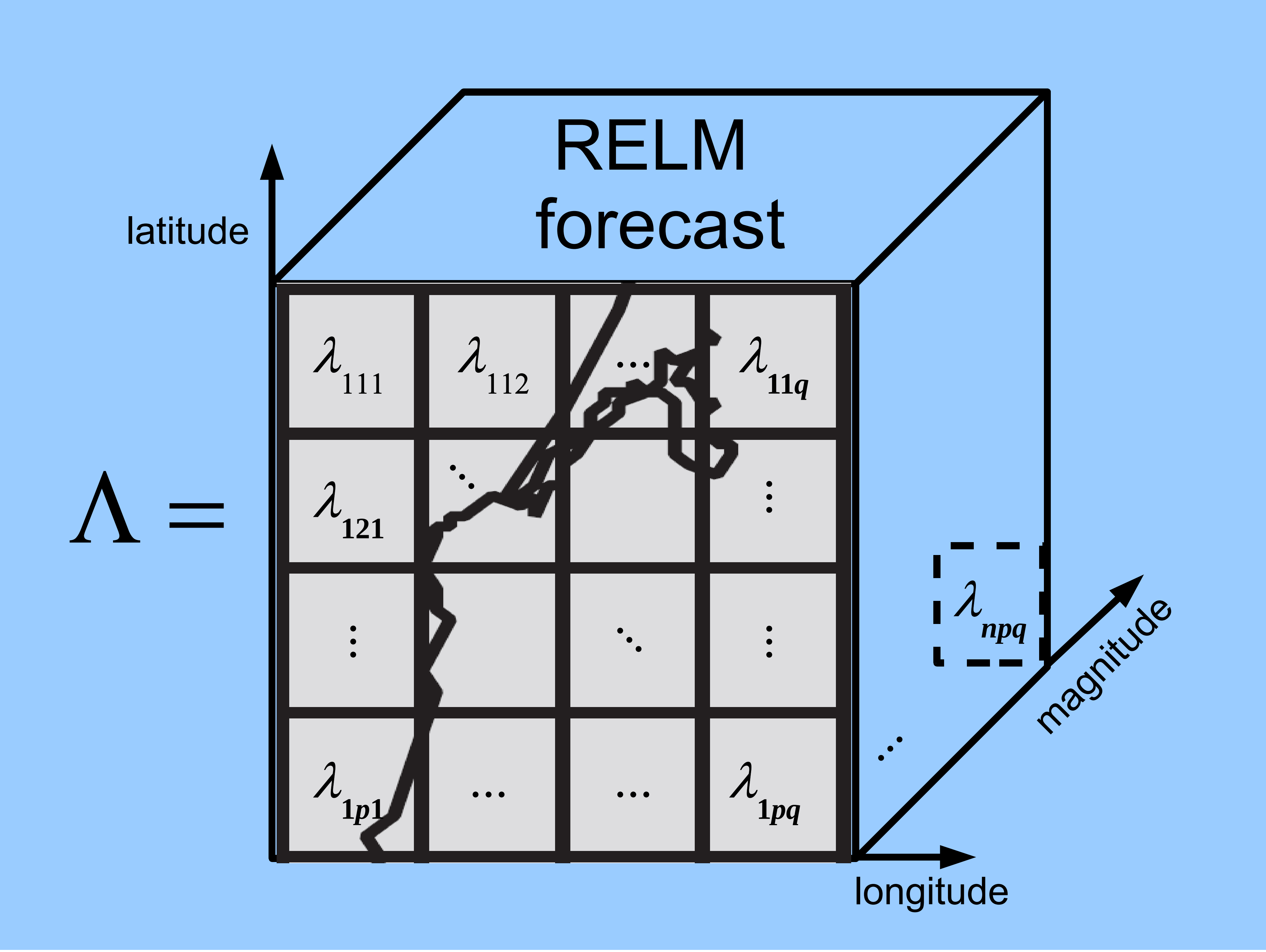
- Likelihood tests, inherited from RELM
- Post-RELM tests
  - Recent extension of likelihood tests
  - Receiver Operating Characteristic (ROC)
  - Molchan error diagram
  - Area skill score
- Process for implementing test

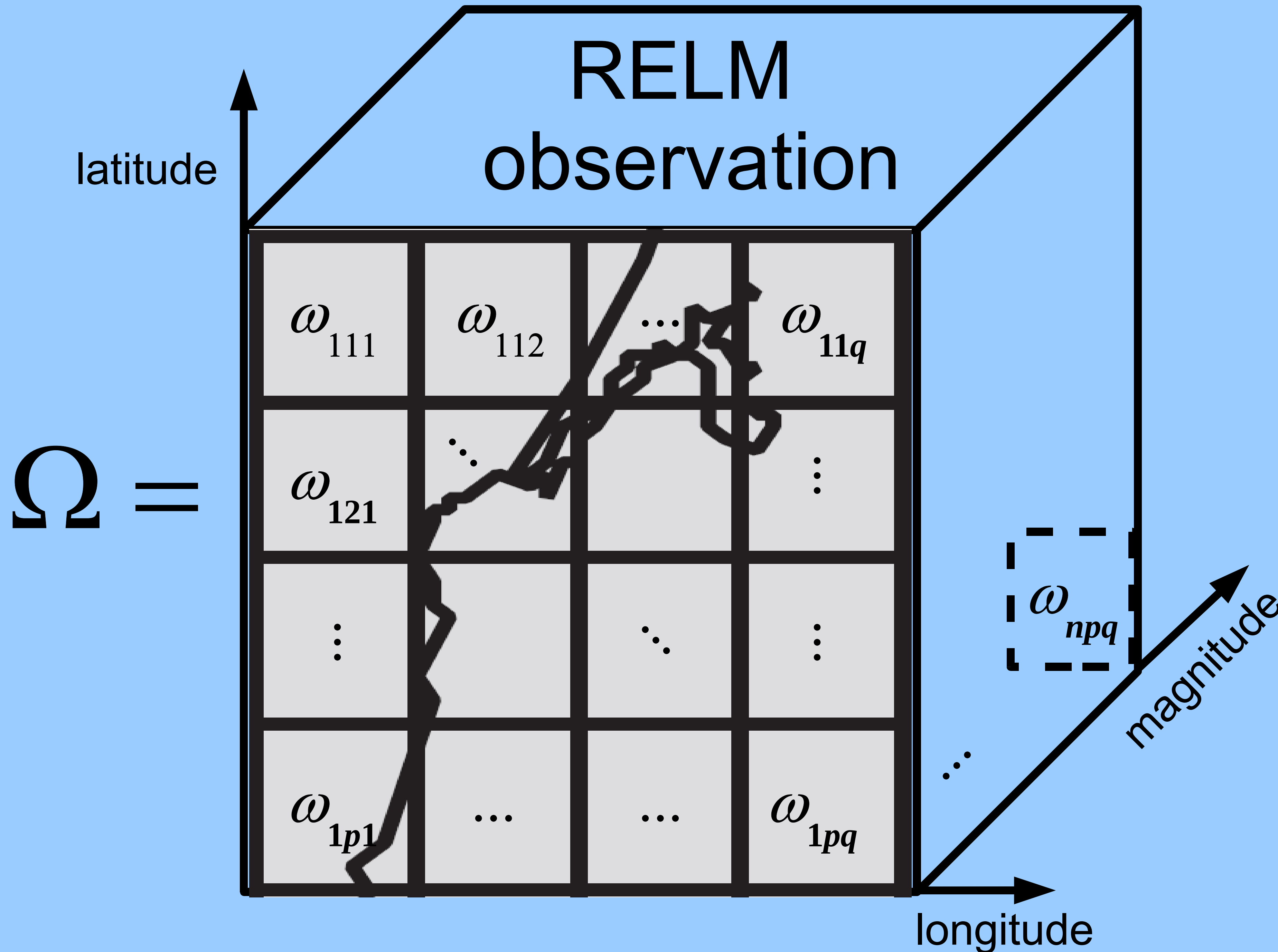
Regional Earthquake Likelihood Model (RELM) experiment in California

- Nineteen *5-year* forecasts
  - Target eqks $M_{ANSS} \geq 4.95$
- Forecasts are specified in terms of expected number of earthquakes in lat/lon/mag bins.
- Currently being tested within CSEP
- Forecasts are evaluated for consistency with observations using likelihood tests.
- Similar format as current prototype testing center in Japan



Schorlemmer & Gerstenberger 2007


RELM
forecast
latitude
longitude
magnitude
$\Lambda =$
λ_{111}
λ_{112}
λ_{11q}
λ_{121}
λ_{1p1}
λ_{1pq}
λ_{npq}



RELM
observation
latitude
$\Omega =$
ω_{111}
ω_{112}
ω_{11q}
ω_{121}
ω_{1p1}
ω_{1pq}
ω_{npq}
magnitude
longitude

RELM likelihood statistics

Number of earthquakes forecast

$$N_{\Lambda}=\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{k=1}^{q}\lambda_{ijk}$$

Log-likelihood of observation given forecast

$$L(\Omega|\Lambda)=\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{k=1}^{q}\left(\omega_{ijk}\log\lambda_{ijk}-\lambda_{ijk}-\log\omega_{ijk}!\right)$$

Log-likelihood ratio of two forecasts

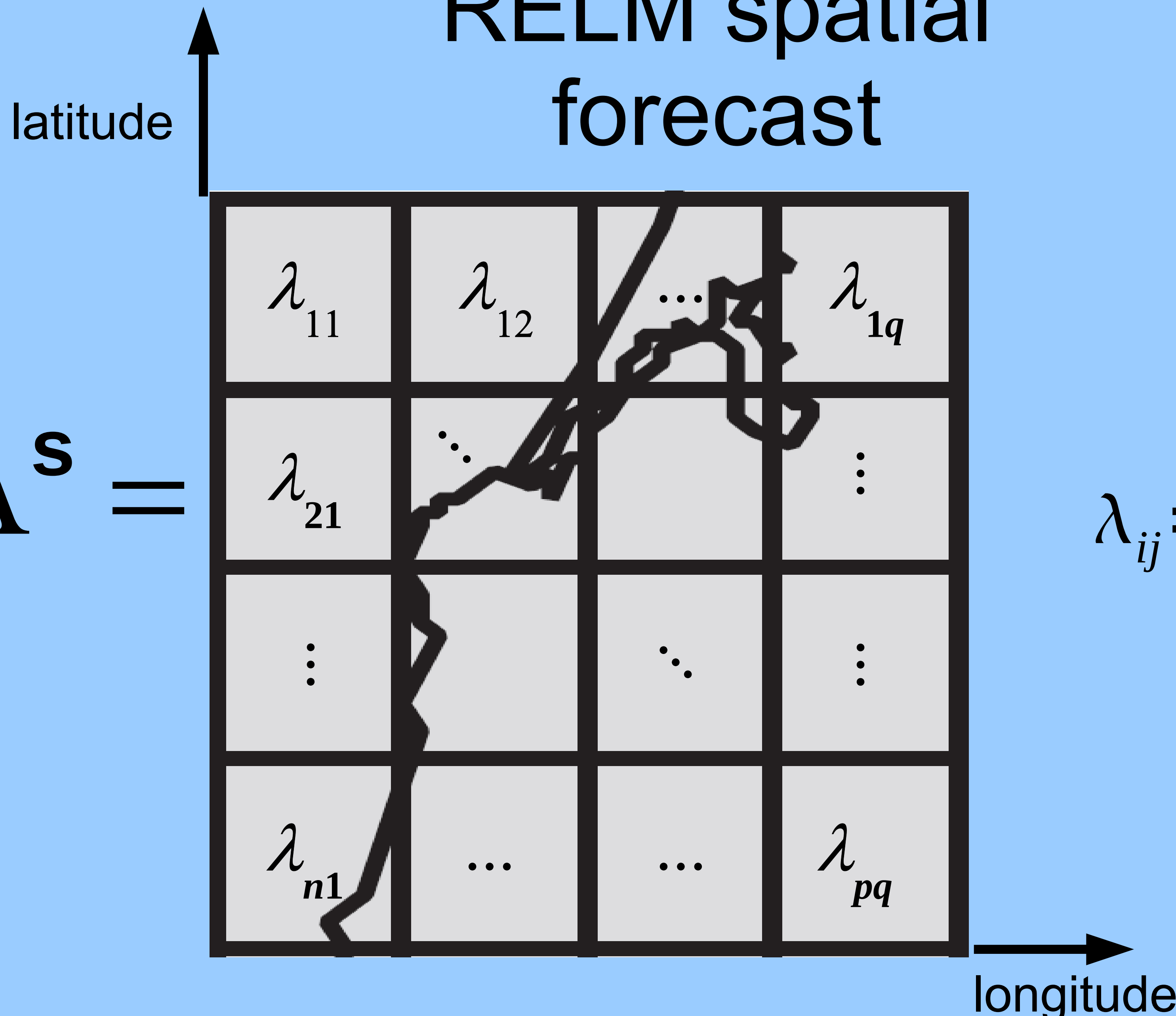
$$R_{xy}=L(\Omega|\Lambda_{x})-L(\Omega|\Lambda_{y})$$

RELM evaluation metrics

- $\underline{\delta}$ Compare number forecast with number observed—did the forecast predict an unreasonably high or unreasonably low seismicity rate?
- $\underline{\gamma}$ Compare forecast distribution with observed distribution—did the forecast obtain an unreasonably low log-likelihood?
- $\underline{\alpha}$ Compare log-likelihood ratio of two forecasts—when taken as null hypothesis, can a forecast be “rejected” by another? And *vice versa*?
- Further details (e.g., catalog uncertainty) in Schorlemmer *et al.* 2007 SRL

RELM spatial forecast

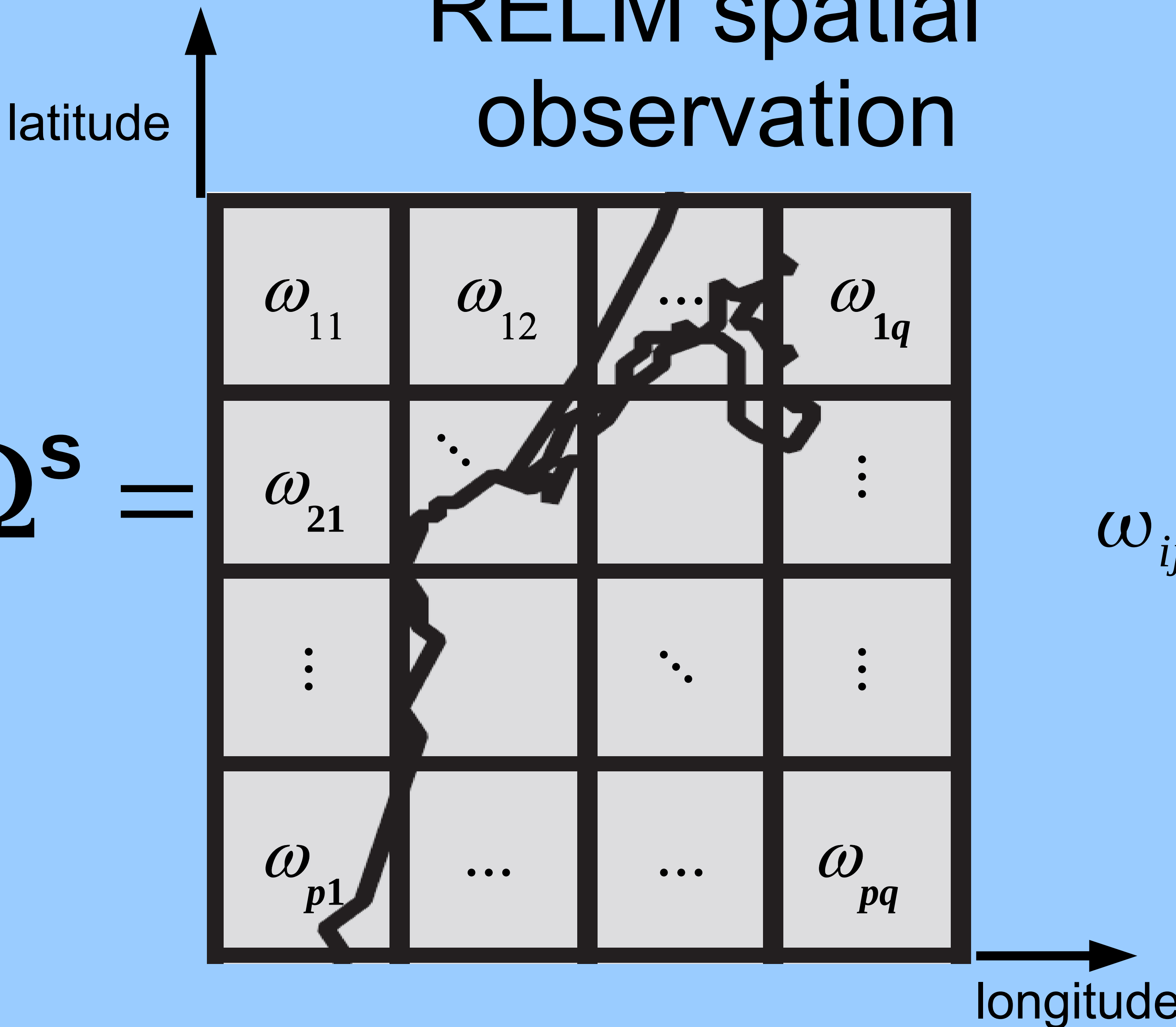
$\Lambda^{s} =$



$$\lambda_{ij} = \sum_{k=1}^{q} \lambda_{ijk}$$

RELM spatial observation

$$\Omega^{s} =$$



$$\omega_{ij} = \sum_{k=1}^{q} \omega_{ijk}$$

Post-RELM likelihood statistics

Log-likelihood of observation given *spatial* forecast

$$S(\Omega^s|\Lambda^s) = \sum_{i=1}^{p} \sum_{j=1}^{q} (\omega_{ij}^s \log \lambda_{ij}^s - \lambda_{ij}^s - \log \omega_{ij}^s!)$$

Log-likelihood of observation given *magnitude* forecast

$$M(\Omega^m|\Lambda^m) = \sum_{i=1}^{n} (\omega_i^m \log \lambda_i^m - \lambda_i^m - \log \omega_i^m!)$$

Post-RELM evaluation metrics

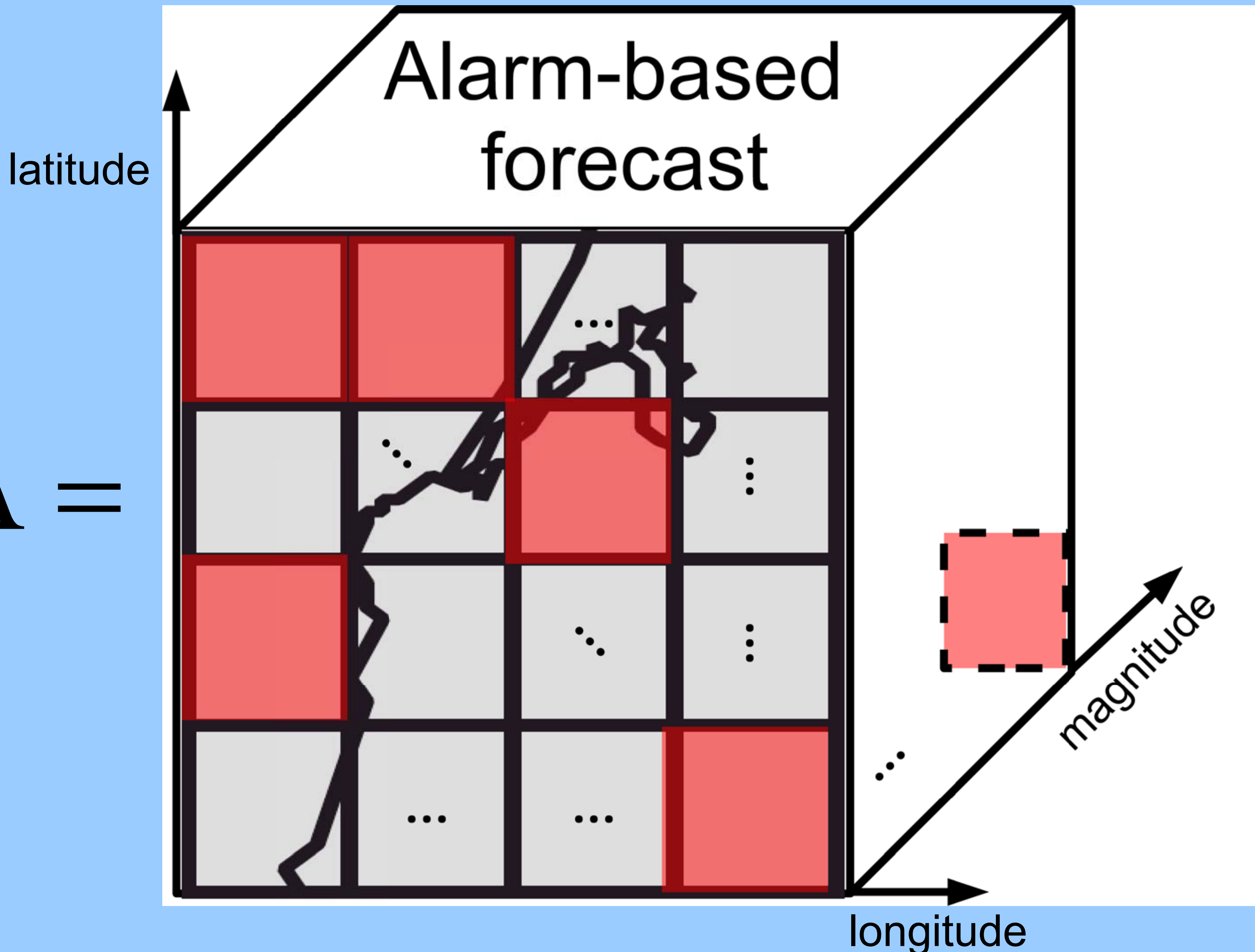
- $\underline{\zeta}$ Compare forecast spatial distribution with observed spatial distribution—did the forecast obtain an unreasonably low log-likelihood?

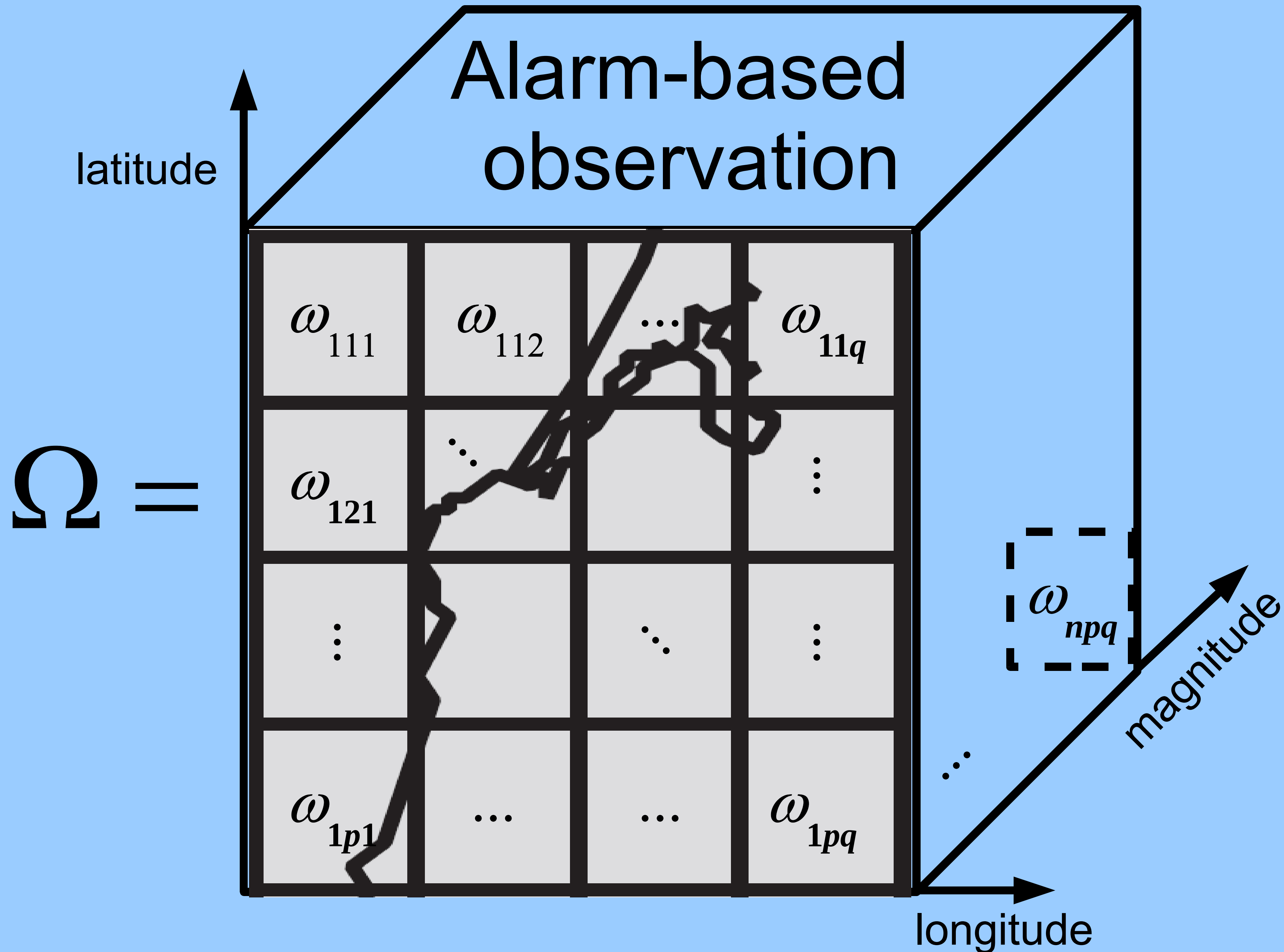
- $\underline{\kappa}$ Compare forecast magnitude distribution with observed magnitude distribution—did the forecast obtain an unreasonably low log-likelihood?

Likelihood tests summary

- Tests require gridded rate forecasts
- Each forecast is characterized by
  - Single δ, γ, ζ, κ value
  - Vector of α values (when comparing *N* forecasts, *N*-1 elements)
- Results presented as plots or tables
  - Temporal variation of δ, γ, α, ζ, κ values
- Implemented in Python/MATLAB

$$\Lambda =$$





Alarm-based observation
latitude
longitude
magnitude
$\Omega =$
ω_{111}
ω_{112}
ω_{11q}
ω_{121}
ω_{1p1}
ω_{1pq}
ω_{npq}

Alarm-based statistics

Hit rate

$$H_\Lambda = \frac{\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{k=1}^{q}\left(\lambda_{ijk}\cdot\omega_{ijk}\right)}{\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{k=1}^{q}\omega_{ijk}}$$

False alarm rate

$$F_\Lambda = \frac{\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{k=1}^{q}\left(\lambda_{ijk}\cdot 1_{\omega_{ijk}=0}\right)}{\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{k=1}^{q}\lambda_{ijk}}$$

Fraction of space occupied by alarm*

$$\tau_\Lambda = \frac{\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{k=1}^{q}\lambda_{ijk}}{n\cdot p\cdot q}$$

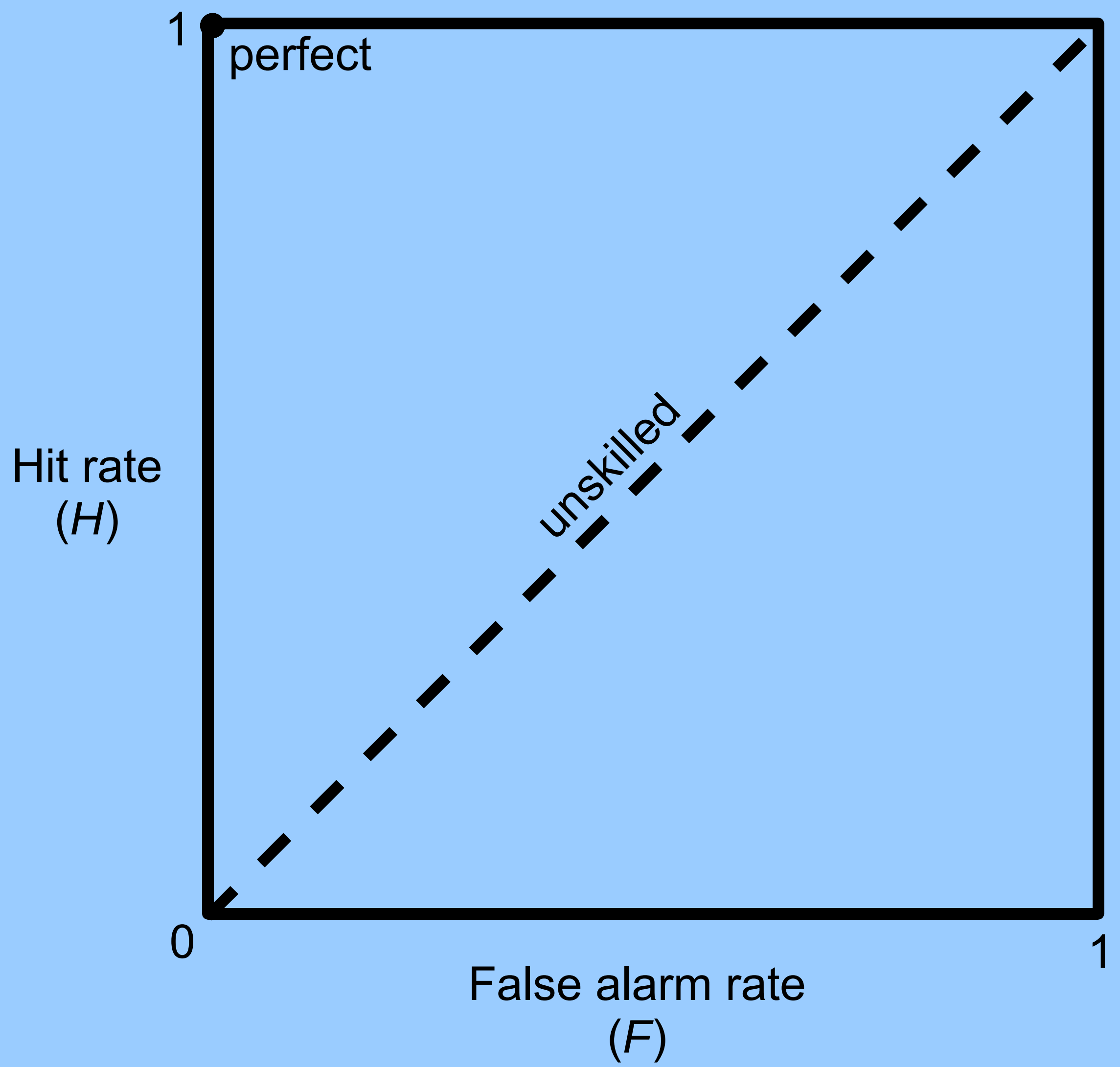
Miss rate

$$\nu_\Lambda = 1 - H_\Lambda$$

Alarm-based evaluation tools/metrics

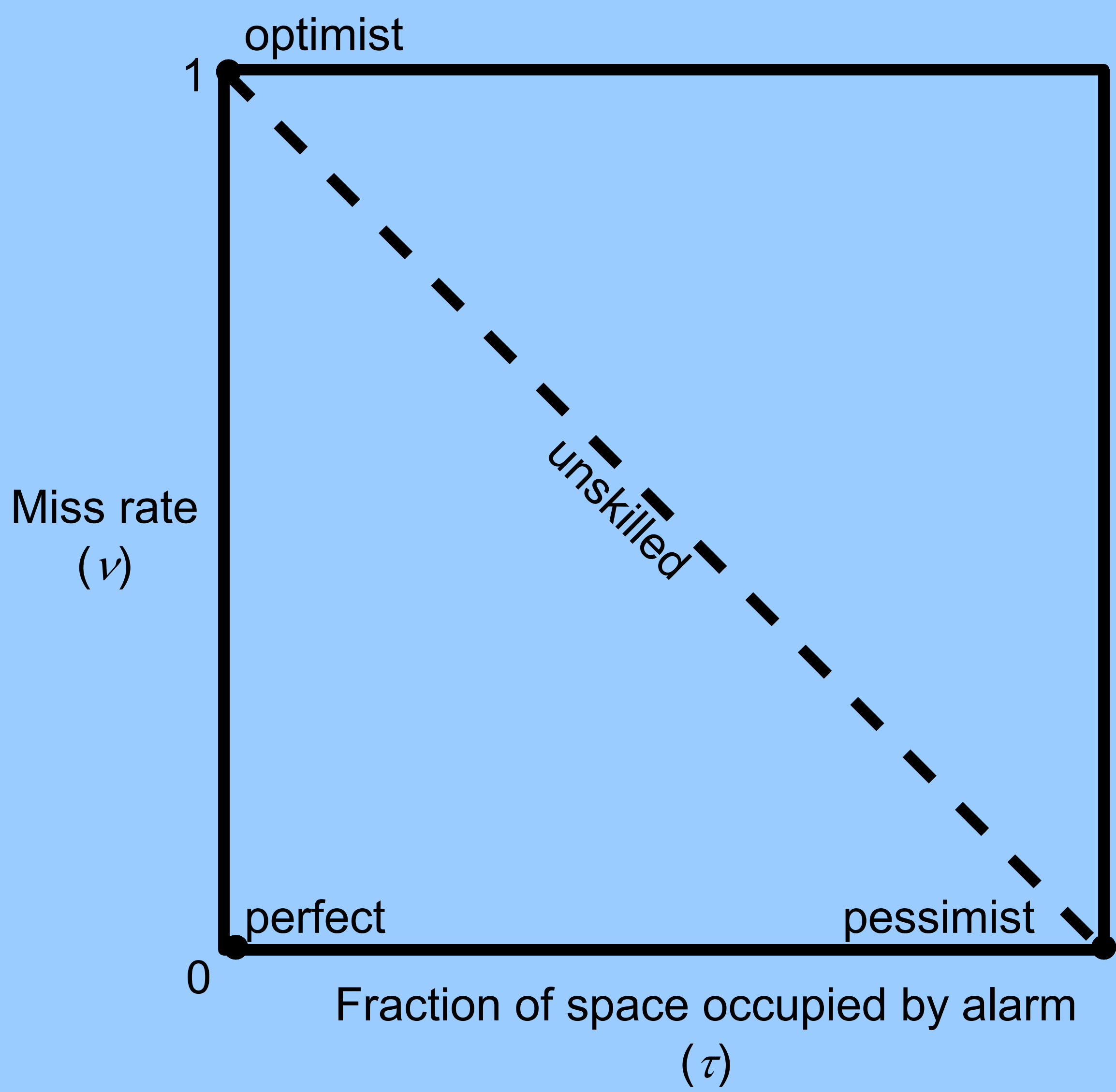
- Receiver Operating Characteristic (ROC) – hit rate and false alarm rate
- Molchan error diagram – fraction of space occupied by alarm and miss rate
- Area skill score – derived from Molchan error diagram

ROC diagram



Mason 2003

Molchan error diagram



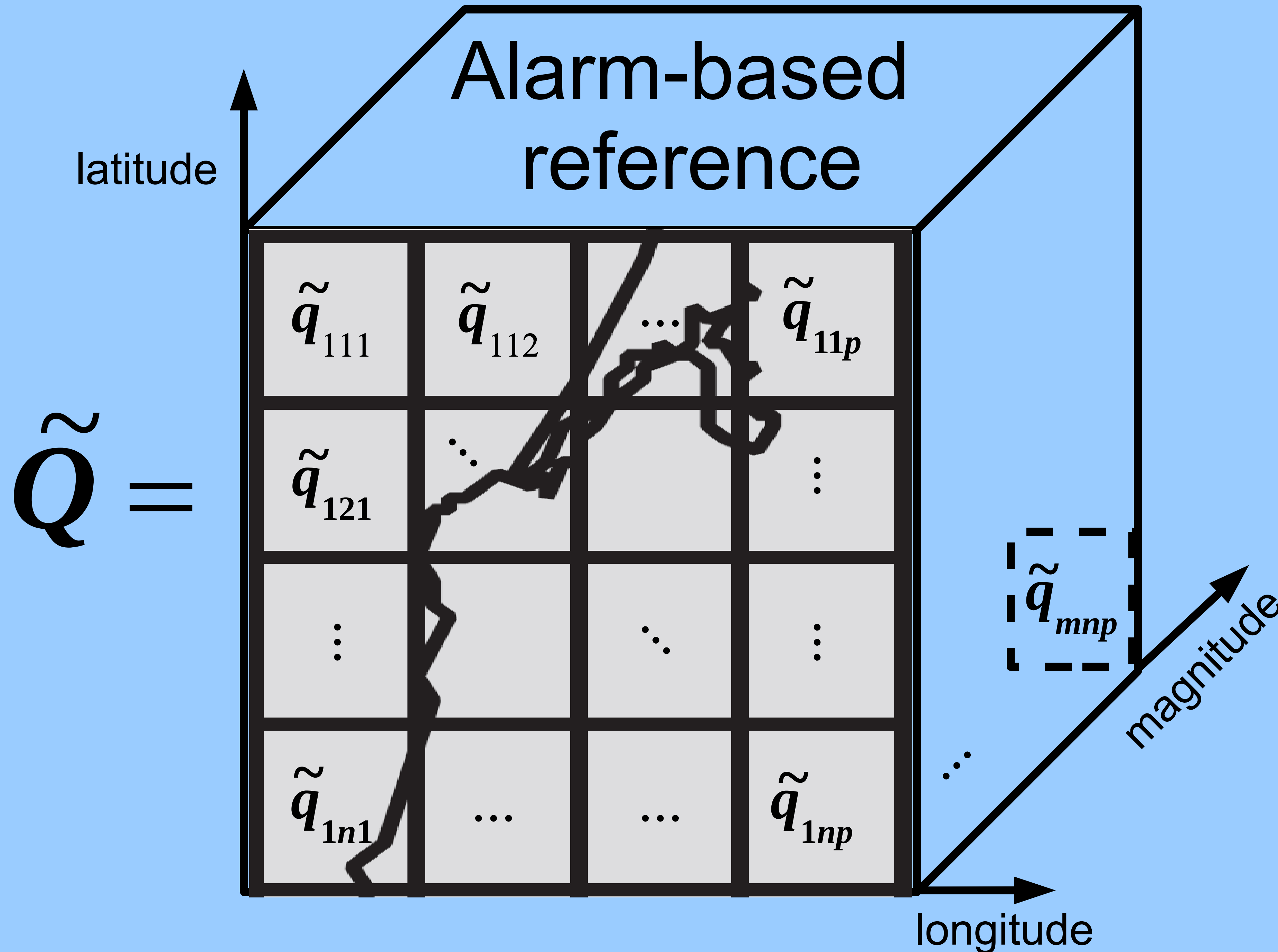
Molchan 1991, Molchan & Kagan 1992

Difference in unskilled reference

- For spatial forecast
  - ROC reference forecast is uniform
  - Molchan reference forecast is "user-defined," should be best estimate of spatial distribution of seismicity

Fraction of space occupied by alarm

$$\tau_{\Lambda} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{k=1}^{q} \left(\lambda_{ijk} \cdot \tilde{q}_{ijk} \right)}{\sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{k=1}^{q} \tilde{q}_{ijk}}$$


Alarm-based reference
latitude
$\tilde{Q} =$
$\tilde{q}_{111}$
$\tilde{q}_{112}$
...
$\tilde{q}_{11p}$
$\tilde{q}_{121}$
$\tilde{q}_{1n1}$
$\tilde{q}_{1np}$
$\tilde{q}_{mnp}$
magnitude
longitude

Post-RELM tests summary

- Tests require alarm function
  - Forecast orders regions of space/time/magnitude
  - Do not require rates
  - In principle, do not require gridding
- Tests allow/require choice of reference forecast
- Each forecast is characterized by
  - ROC: vector of (*H*, *F*) values (*N* elements, *N* is number of observed eqks)
  - Molchan: vector of (τ, ν) values (*N* elements)
  - Single area skill score value ($a_{\Lambda}(\tau$=1))
- Results presented as plots

Requirements for new test

- Scientific justification
- Existence or introduction of suitable forecasts
- Technical
  - Codes should accept ForecastML and ZMAP formats as input, output ResultsML format
  - Codes should be documented and software dependencies stated explicitly

Process for new test implementation

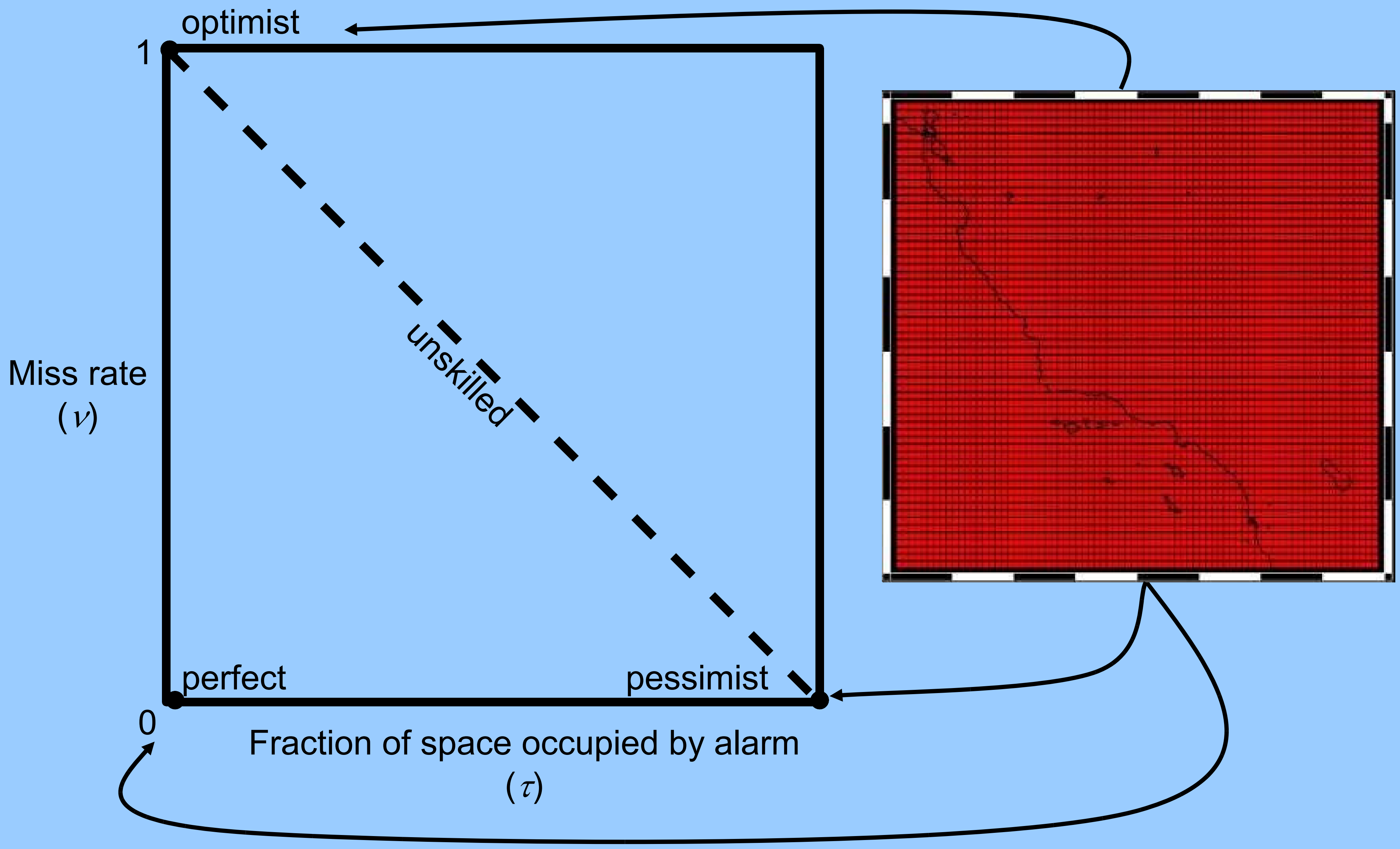
- Work with CSEP Testing Center development team
  - Provide testing codes and support documentation
  - Provide reference data for unit test
    - For a given forecast, what is the expected result?
  - Use of random numbers constitutes a special case

- Aim for updates to operational system on quarterly basis

- Further details on CSEP computational infrastructure in Zechar *et al.* (ms. in review)

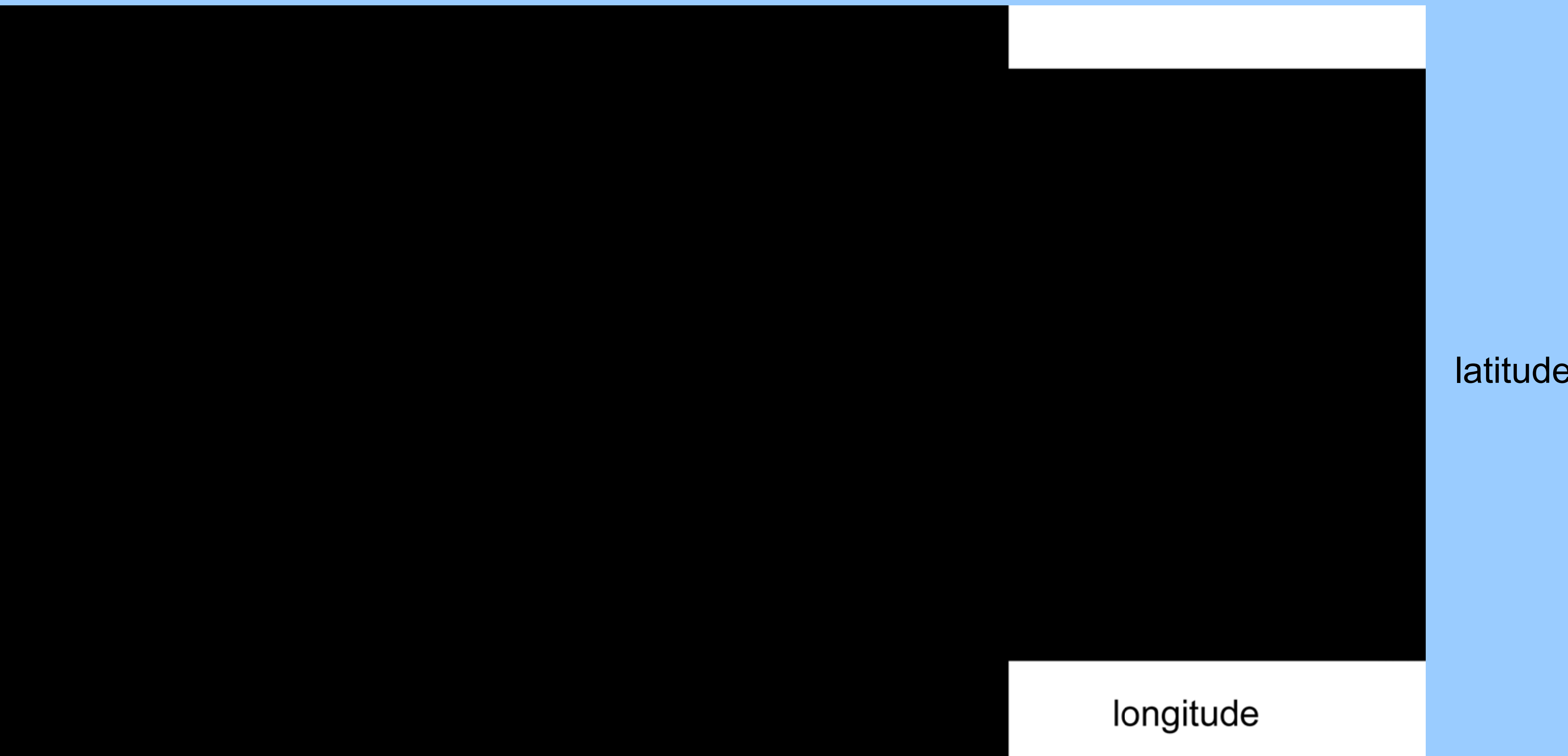
Thank you.

Molchan error diagram

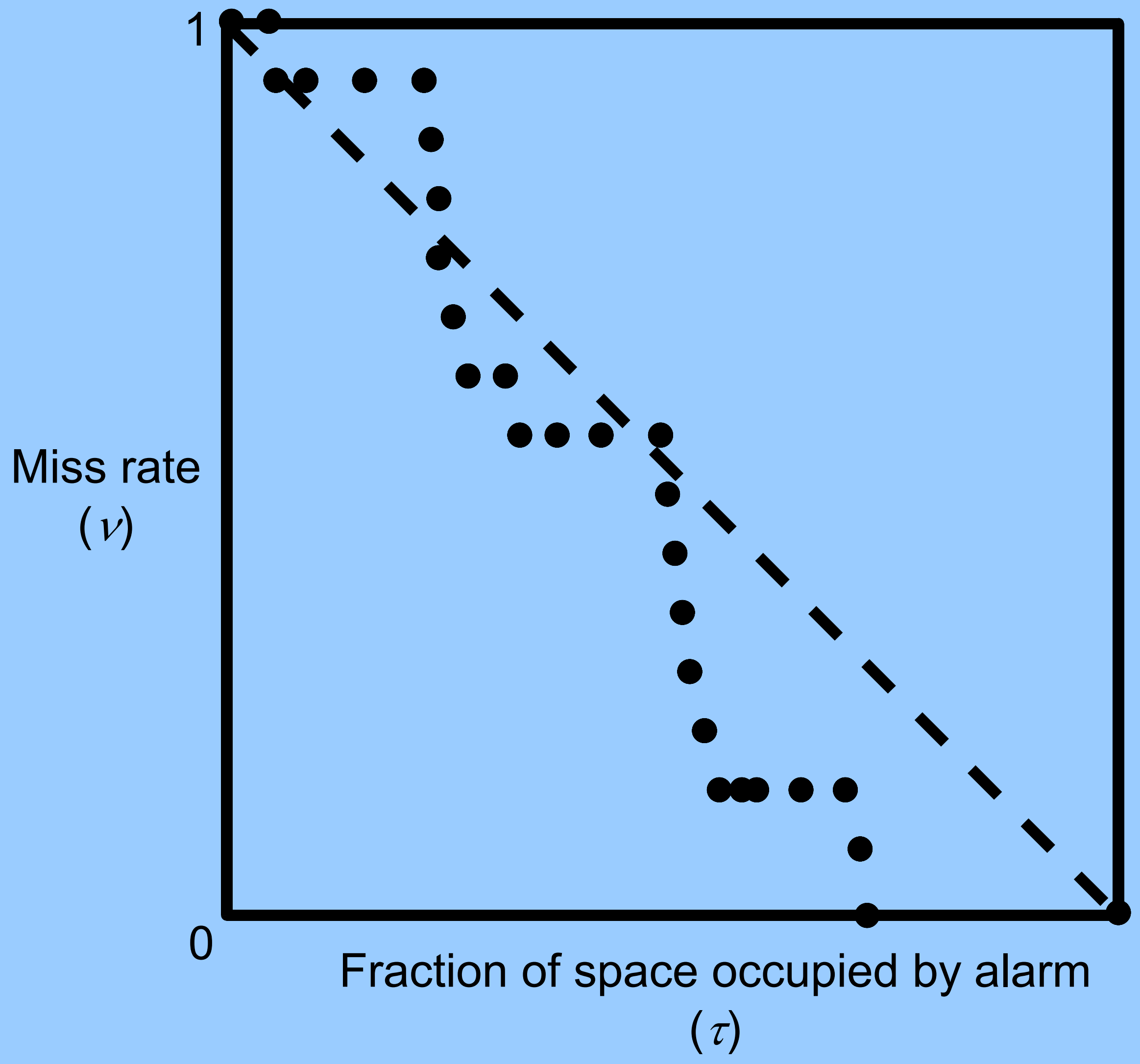


Molchan 1991, Molchan & Kagan 1992

Generalize alarm set to alarm function



Molchan diagram



Molchan trajectory: collection of (τ, ν) points generated from alarm function

Area skill score

- Area above Molchan trajectory, normalized by τ

$$a_\Lambda(\tau) = \frac{1}{\tau}\int_0^\tau \left(1 - \nu_\Lambda(t)\right) dt$$

- Unskilled forecasts yield area skill score ~ ½; forecast skill is characterized by deviation.
- Further details in Zechar & Jordan 2008 GJI, Zechar & Jordan 2009 PAGEOPH