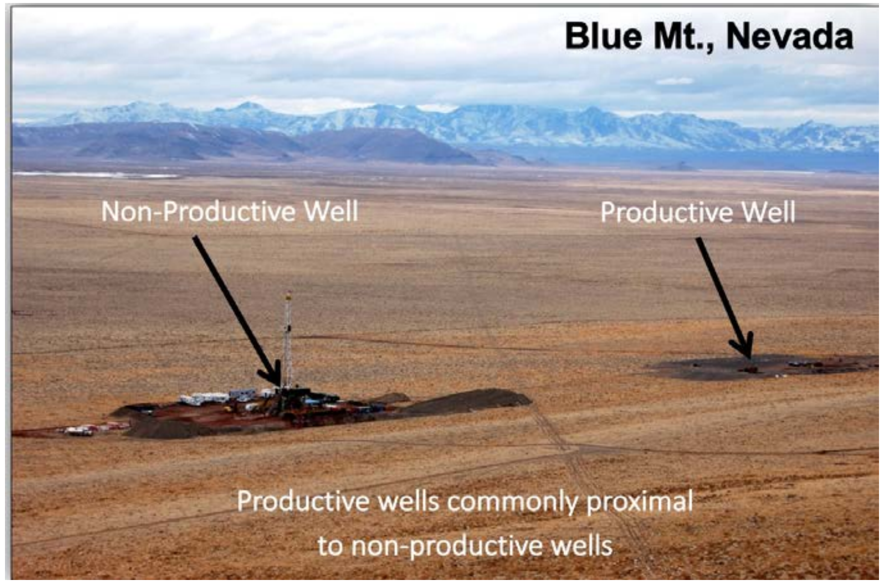
Machine Learning for Identifying Subsurface Targets Using Small Datasets: Applications to Geothermal Energy and Carbon Sequestration

Michael Fehler

Senior Research Scientist, EAPS

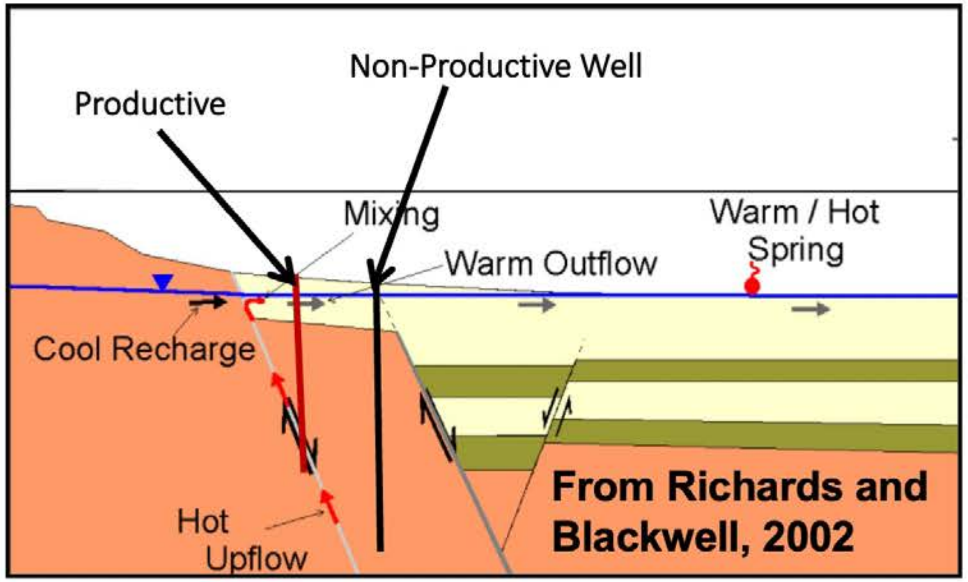
In collaboration with Steve Brown, Bill Rodi, M. Nafi Toksöz, and Chen Gu

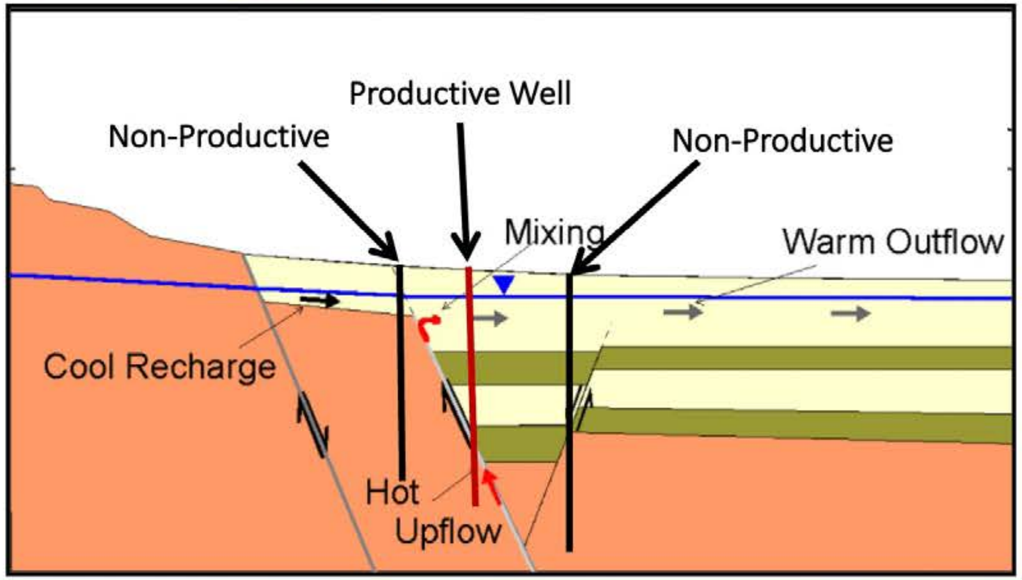
May 26, 2022

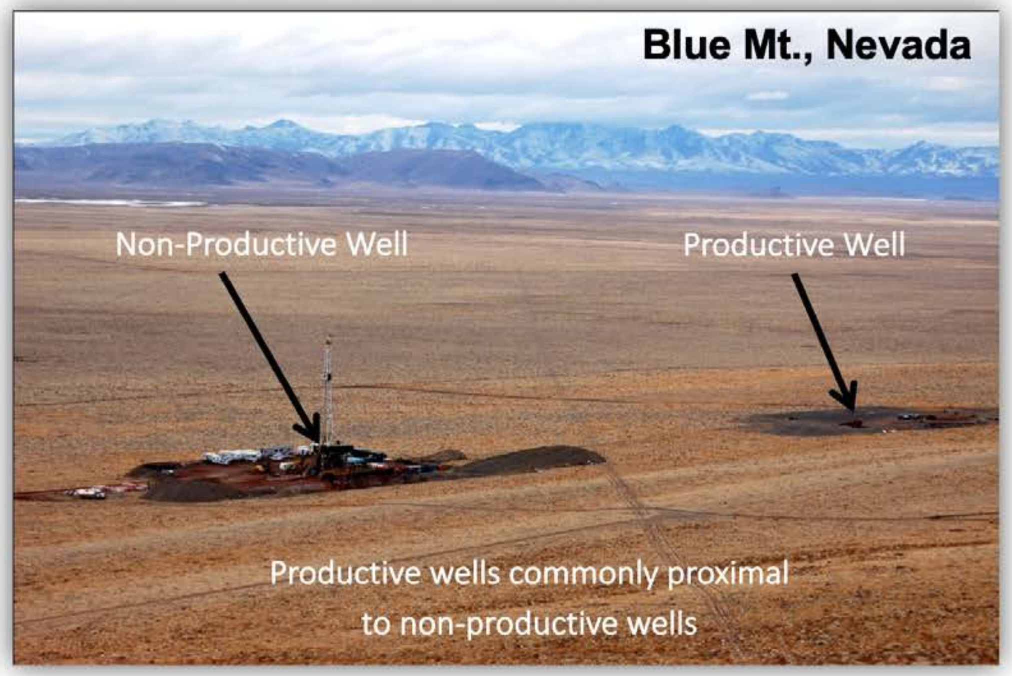


Assistance in conducting geothermal exploration work from Jim Faulds, Sven Treitel, John Queen and Connor Smith

Major Challenge – Hidden Geothermal Systems



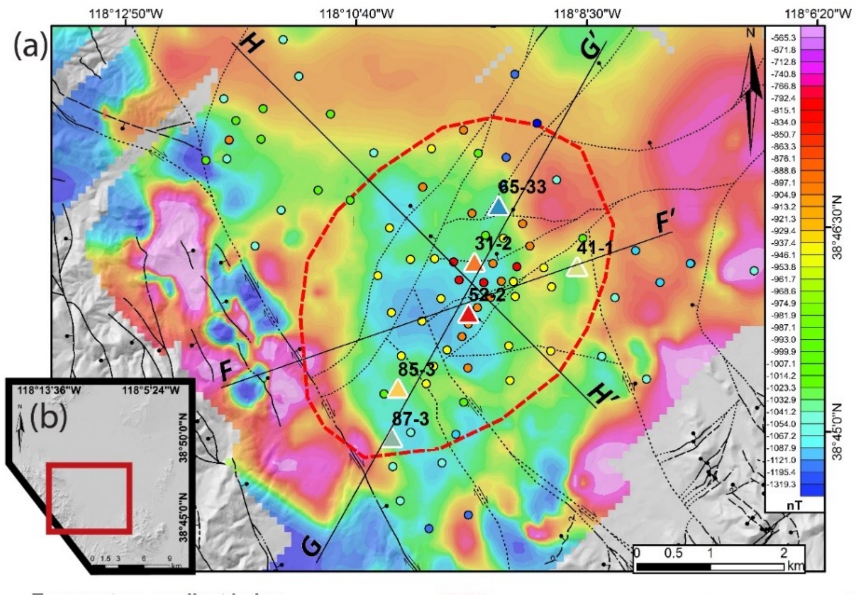




- **40% of known systems blind**
- **Estimated 75% of all systems hidden**
- **Significant drilling risk**
- **Need high-quality geologic and geophysical data to elucidate surface and subsurface to permit development of better conceptual models**

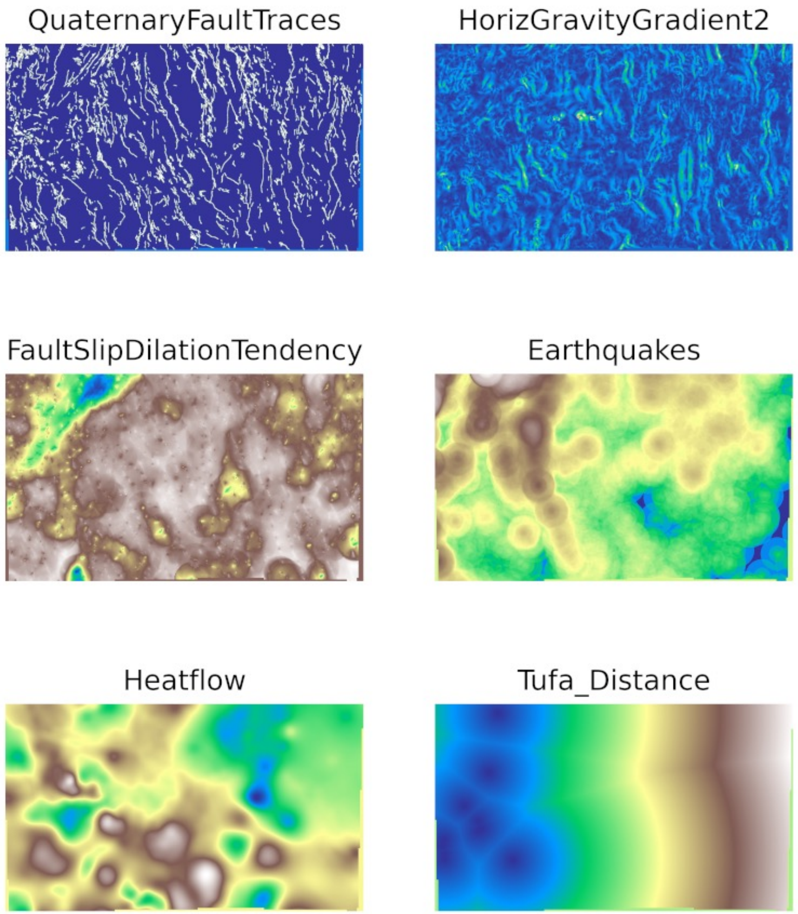
Challenges in Applying Machine Learning to Geothermal Exploration

- **Some geophysical techniques do not work or are too expensive**
  - **e.g. Active seismic**
- **Few training sites**
  - **Can lead to overfitting**
  - **Need both positive and negative sites**
- **Mix of data types:**
  - **Numerical variables (temp, distance to fault, gravity values)**
  - **Categorical variables (mineral assemblage, rock type)**
  - **Ordinal variables (size of feature)**
  - **Variation in resolution or uncertainty**
- **Some features not continuous**
  - **Requires special treatment to avoid bias**

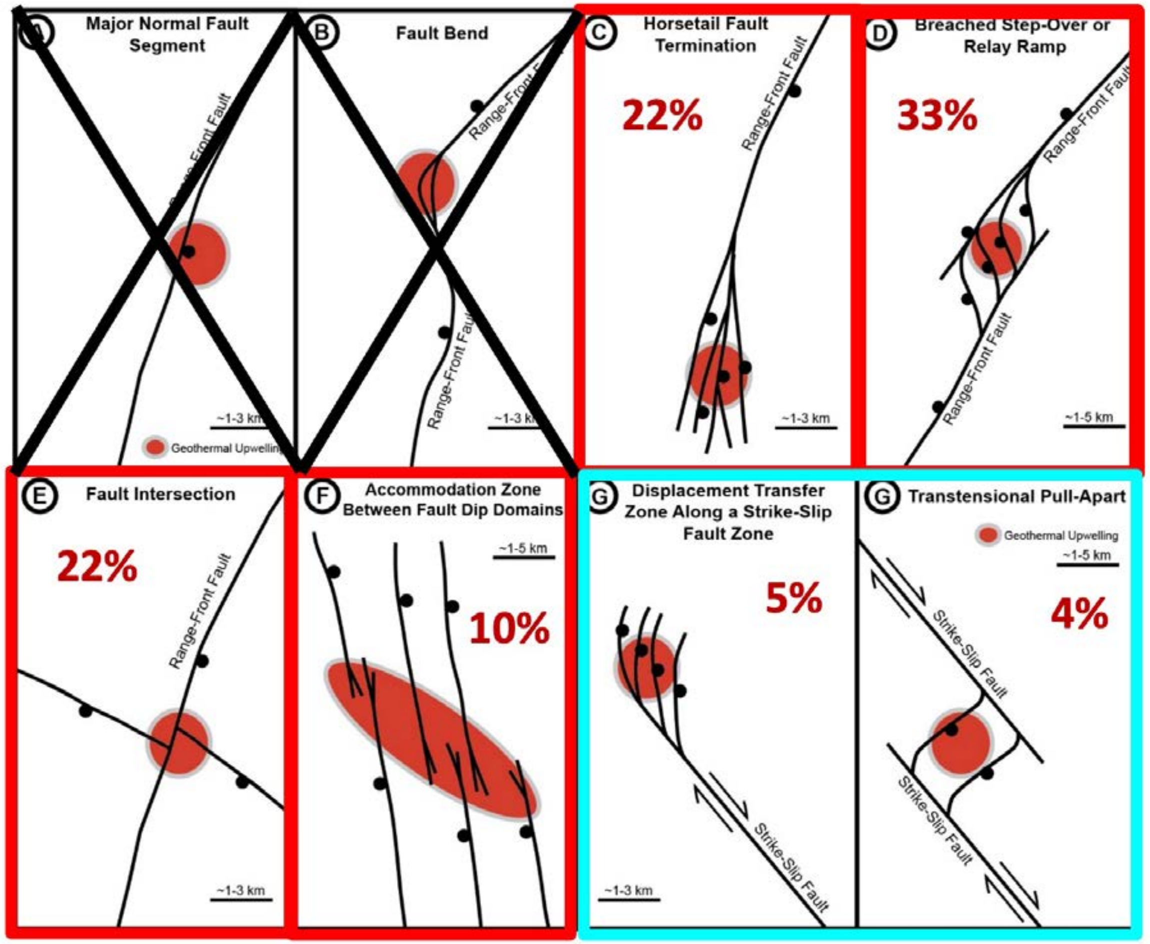


Example Numerical and Categorical Features for Nevada Study Area

Numerical



Categorical: Structural Setting



Site Classification Problem

- A potential geothermal site is associated with
  - A class y:

$$y = \begin{cases} 1, \text{if the site is a positive prospect} \\ 0, \text{if the site is a negative prospect} \end{cases}$$

  - A feature vector $\boldsymbol{x} = (x_1, x_2 \dots \dots)$ containing geological, geophysical and geochemical measurements/information
- The **Prediction** Problem
  - Given $\boldsymbol{x}$, what is the probability that the site is $y = 1$ *vs.* $y = 0$
  - e.g. infer $\mathrm{Prob}\{y|\boldsymbol{x}\}$ (and its uncertainty)
- The **Decision** problem
  - Given $\mathrm{Prob}\{y|\boldsymbol{x}\}$ and its uncertainty, decide if $y = 1$ or $y = 0$

    (considering mis-calculation risks, etc.)
- We focus on the **Prediction** problem

Fully-Connected Network

Parameterize the probability of a positive site with a multilayer neural network, P:

$$p \stackrel{\text{def}}{=} \text{Prob}(y = 1|\boldsymbol{x}) = P(\boldsymbol{x}, \boldsymbol{\gamma})$$

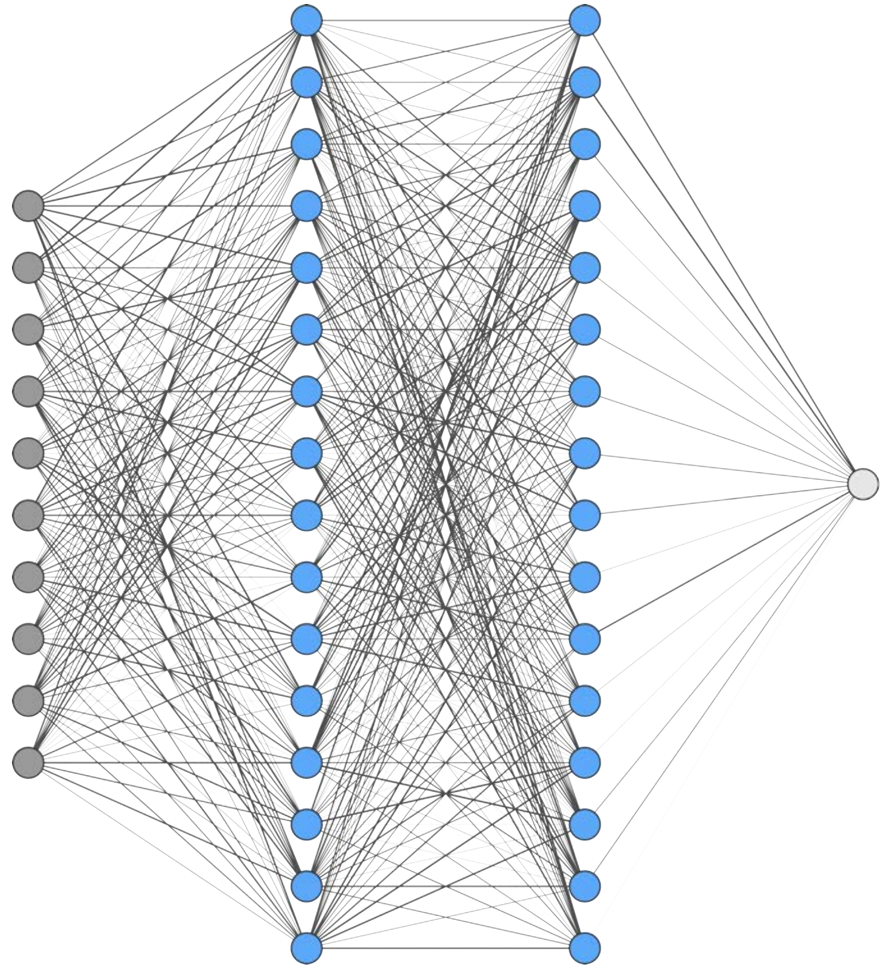
where

$\boldsymbol{x}$ = Neural Network input

p = Neural Network output

$\boldsymbol{\gamma} \stackrel{\text{def}}{=} (\boldsymbol{w}, \boldsymbol{b})$ = Neural Network weights and biases

Regularize by limiting size of weights and/or biases to limit overfitting and prevent small number of nodes from dominating



- Dark gray on left are 10 input features
- Light gray circle on right is output
- Blue are 16 neurons in each of 2 hidden layers (biases)
- Gray lines are connections (weights)

Traditional Machine Learning Approach

- The prediction problem reduces to inferring $\boldsymbol{\gamma}$ using training data from sites with known class

$$D = \{(x_i \, , y_i), i = 1,2 \ldots \ldots N\}$$

- A traditional (deterministic) Neural Network finds a single, optimal value of $\boldsymbol{\gamma}$, e.g. the *maximum likelihood* estimate:

$$\boldsymbol{\gamma}_{ML} \; = \arg max_{\gamma} \mathcal{L}(D; \boldsymbol{\gamma})$$

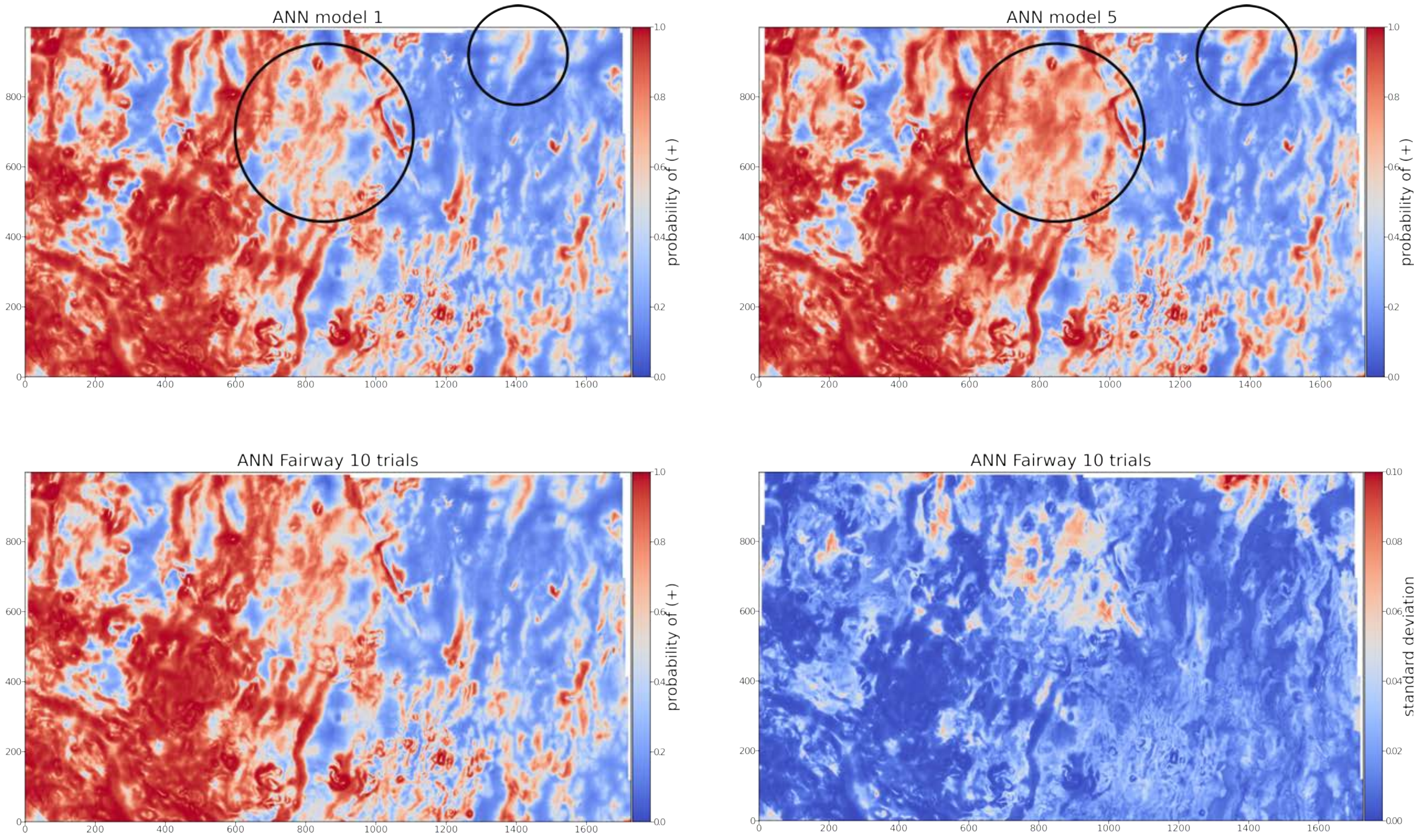
- The Trained Neural Network calculates, for any $\boldsymbol{x}$,

$$p = P(\boldsymbol{x}, \boldsymbol{\gamma}_{ML})$$

- Our algorithm for likelihood maximation initializes $\boldsymbol{\gamma}$ to a random seed

Problem: Variability with Multiple Runs

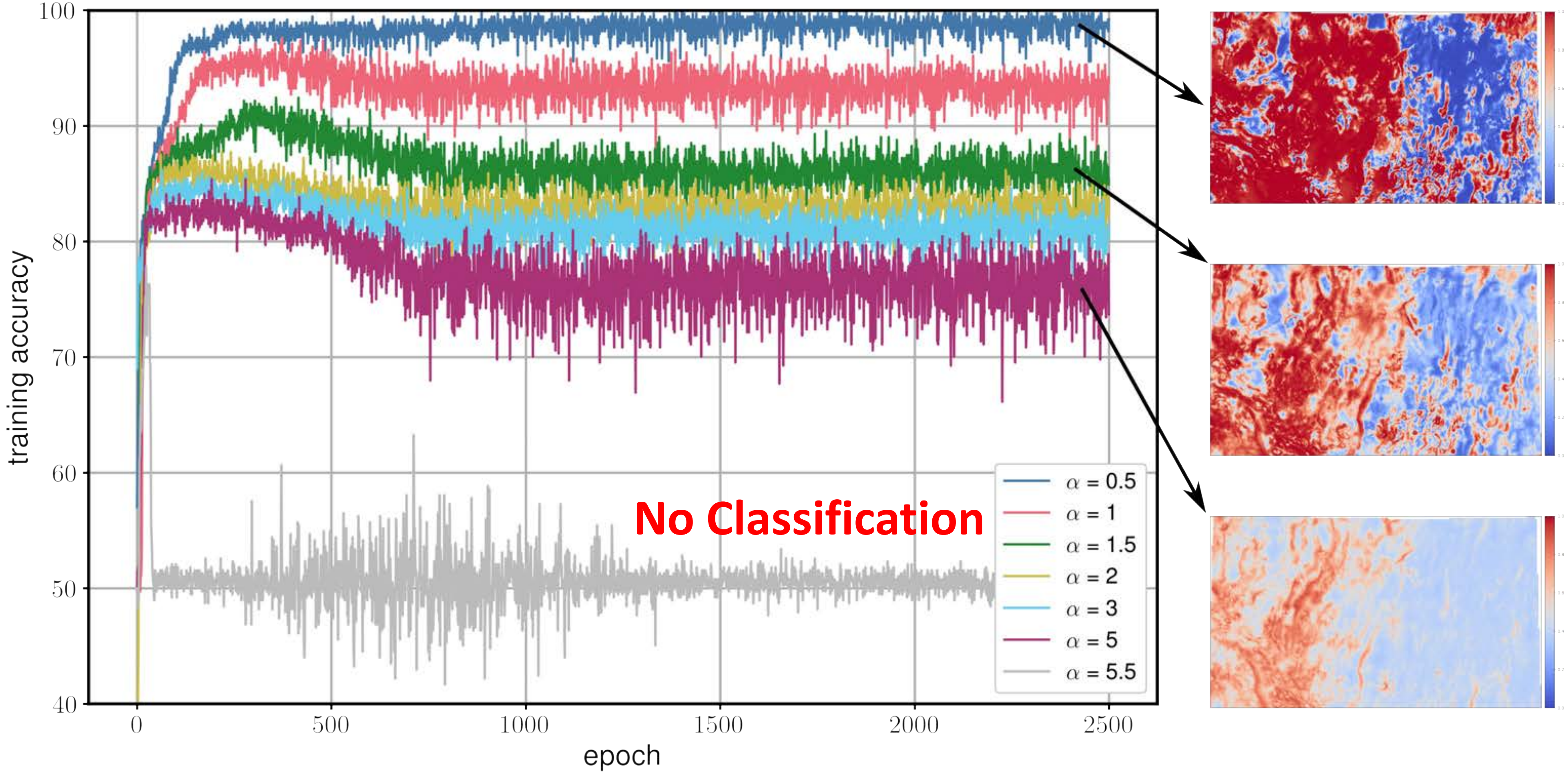
- Ran 10 test models with differing seeds
- Top two images: models 1 and 5 represent extreme variability
- Lower left: mean of 10 runs
- Lower right: standard deviation of 10 runs



Bayesian Approach

- A Bayesian Neural Network applies Bayesian inference with the training data to determine a posterior probability distribution on $\boldsymbol{\gamma}$
- Our training algorithm uses the Variational Bayes method to find an approximation to the posterior distribution on $\boldsymbol{\gamma}$
- Tradeoff model complexity with fitting of training data
  - controlled by regularization parameter α chosen based on estimate of number of *degrees of freedom*
- We sample the distribution of $\boldsymbol{\gamma}$ in the network to get a posterior probability distribution on p

Bayes Approach: Training Curves

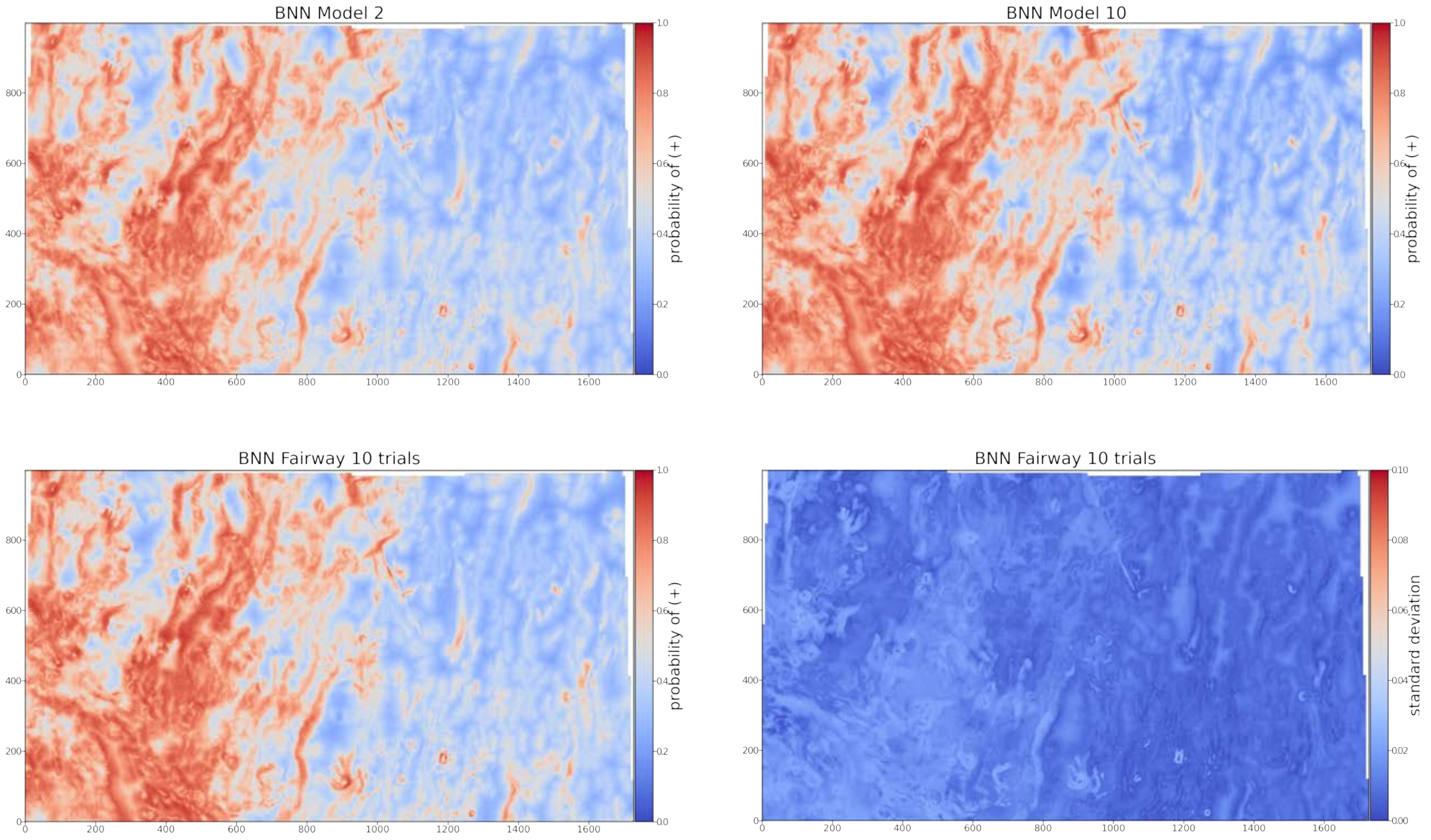


Overfit Data

Increasing NN model complexity

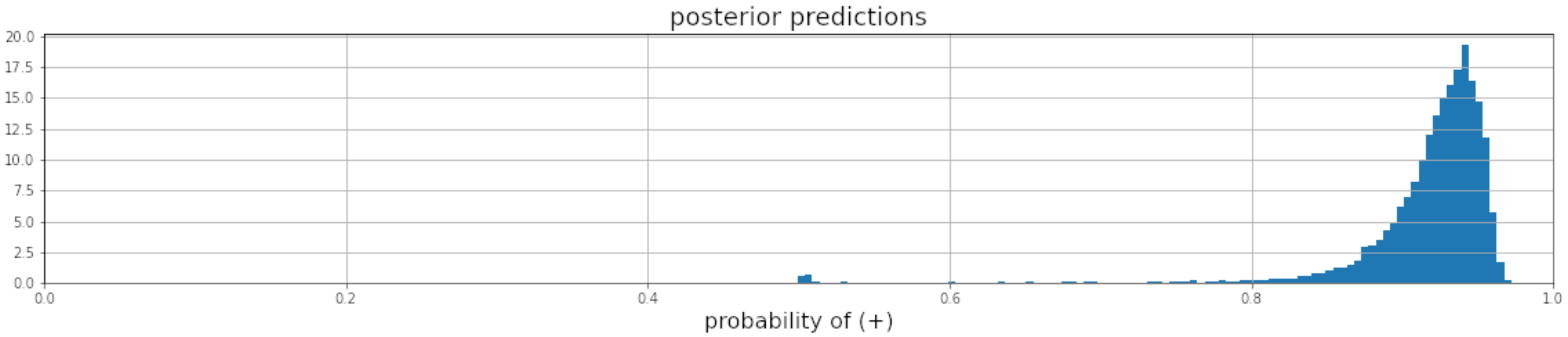
Results Using Bayesian Network

- Ran 10 test models with differing seeds
- Top two images: Maximum variability among multiple runs
- Lower left: mean of 10 runs
- Lower right: standard deviation of 10 runs

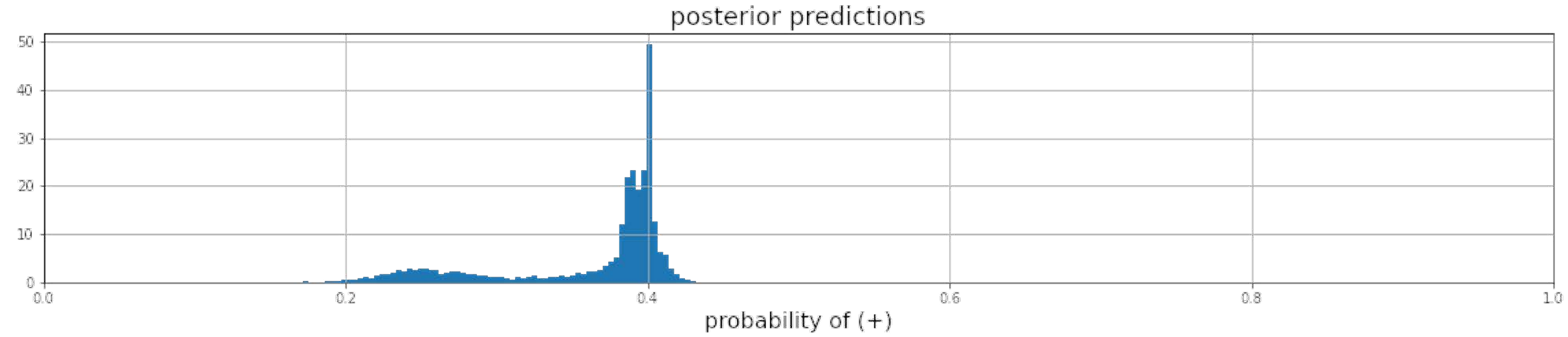


Posterior Probabilities from Bayesian Approach

R588C489 Dixie Meadows (in CNSB blue)

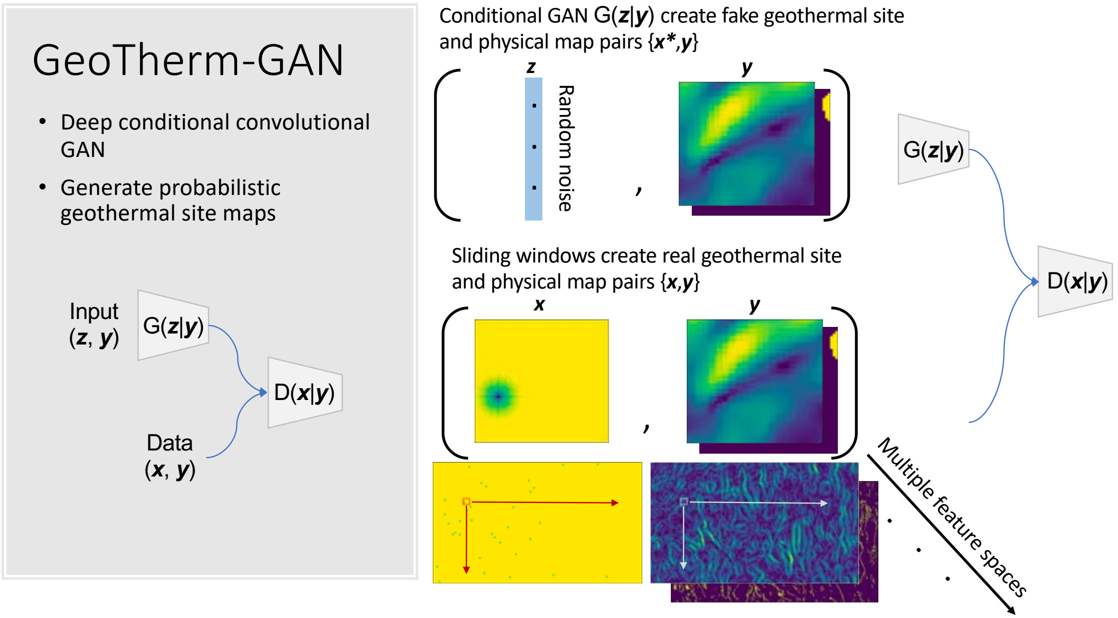


R428C1424 Negative (in CA purple)



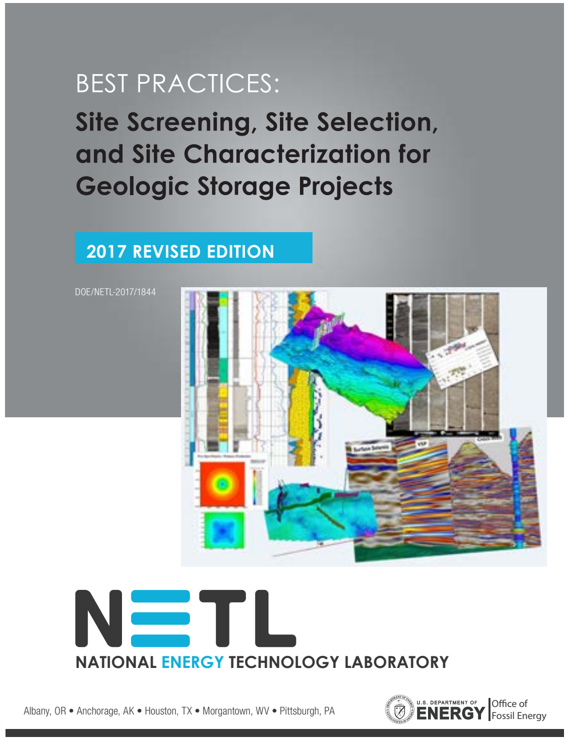
Follow-On Work

- Differing approaches for geothermal
  - Currently use positive and negative geothermal sites
    - Modify positive to be more general – e.g. temperature at depth; production rate, reservoir lifetime
  - Convolutional Neural Network
    - Take advantage of data spatial information
  - Find sites that are most like known geothermal sites (work underway)
  - Develop approach for using categorical data
  - Account for uncertainty in input data
  - Use GAN



Follow-On Work

- Application to site selection for CO_2 sequestration
  - Required characteristics
    - Trapping mechanism
    - Caprock integrity
    - Seismic risk
    - Reservoir capacity
    - Maximum safe injection rate
    - Risk to water table
    - Source of CO_2



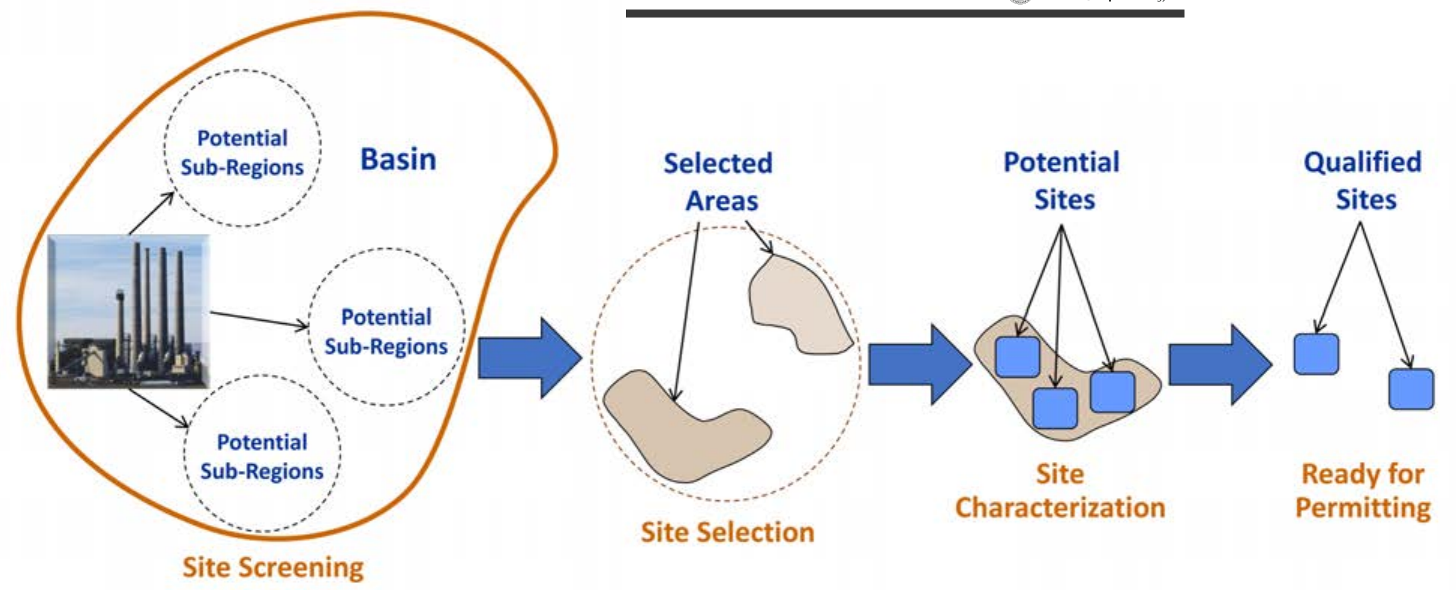
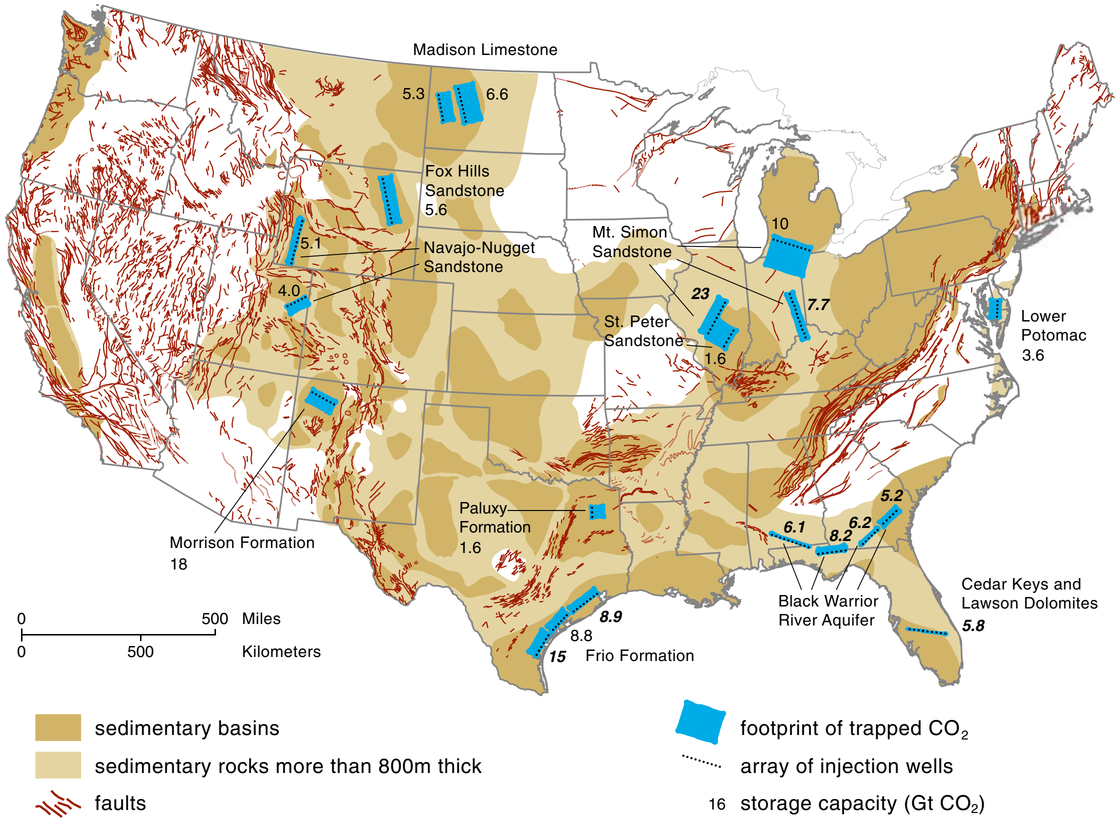


Figure 1.2: Illustration of the Relationship Between Scale of Investigation and Major Steps in Process of Finding and Developing Qualified Sites

Investigate sites for CO_2 Storage

- Start with existing data
  - USGS datasets of regional geophysics
  - Geothermal databases
  - Databases compiled for CO_2 sites
    - Regional partnerships
    - DOE-sponsored studies
  - US Array tomography studies
- Sites identified and tested
  - Regional partnerships
  - Those identified by industry
  - Include EOR sites?

CO_2 Storage Capacity of 20 Sites in US



Michael L. Szulczewski, Christopher W. MacMinn, Howard J. Herzog, and Ruben Juanes, Lifetime of carbon capture and storage as a climate-change mitigation technology, *PNAS,* www.pnas.org/cgi/doi/10.1073/pnas.1115347109, 2012

Framework for Additional Work

- Current geothermal work is near completion
  - Paper(s) being prepared
- Continue geothermal work with industrial partner(s)
  - Explore enhancements to current methodology
  - Apply to different regions where data are available
- Initiate CO_2 work with industrial partner(s) through ERL

Bayesian Approach

- A Bayesian Neural Network applies Bayesian inference to determine a posterior probability distribution on $\boldsymbol{\gamma}$:

$$f(\boldsymbol{\gamma}|D) = \text{"Bayes Rule"}$$

- For any $\boldsymbol{x}$, the Bayesian Neural Network transforms $f(\boldsymbol{\gamma}|D)$ to a probability distribution on p:

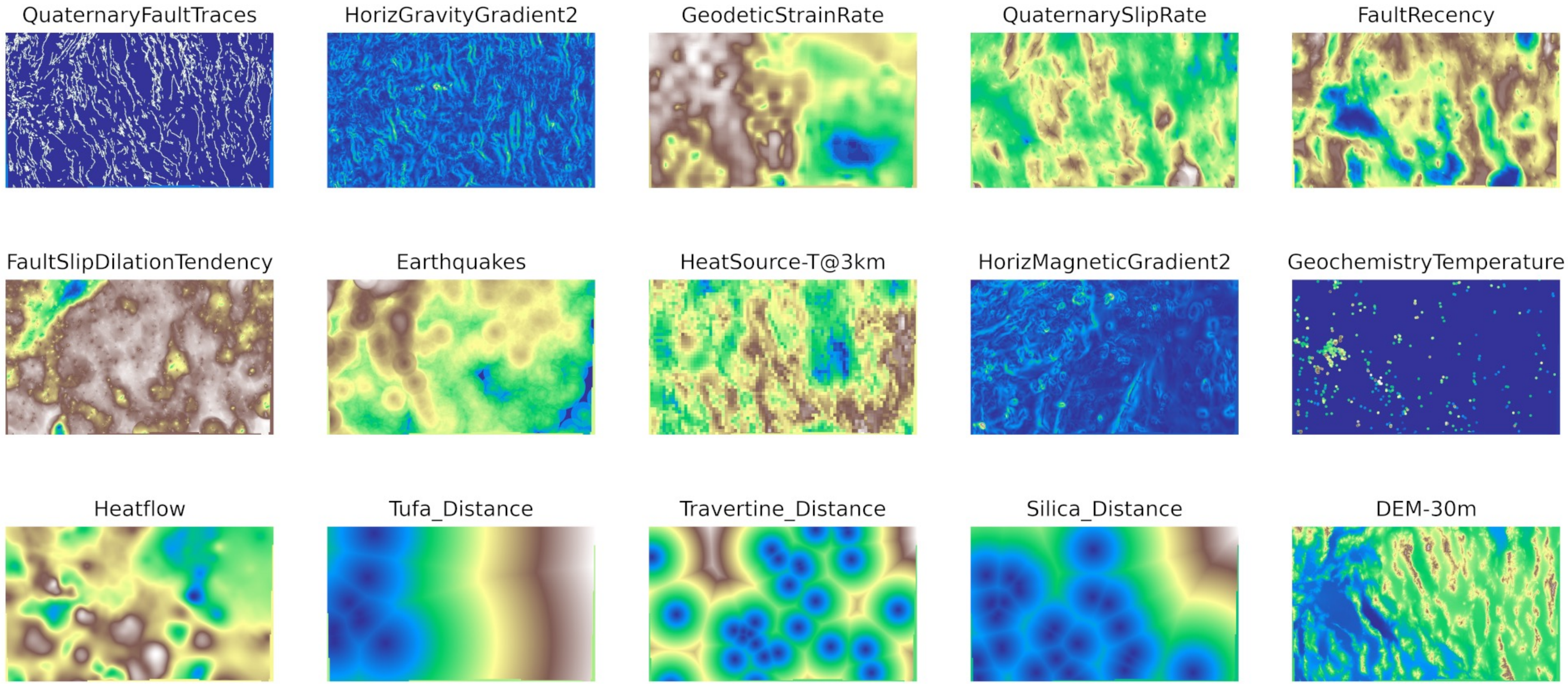
$$f(p|\boldsymbol{x}) = \text{"transformation of } f(\boldsymbol{\gamma}|D)\text{"}$$

Tradeoff model complexity with fitting of training data using regularization parameter chosen based on estimate of number of *degrees of freedom*

$$\mathcal{L}(D, \boldsymbol{\gamma}) = \alpha KL[\mathrm{q}(\boldsymbol{\gamma})||\mathcal{P}(\boldsymbol{\gamma})] + E$$

Where $KL[\mathrm{q}(\boldsymbol{\gamma})||\mathcal{P}(\boldsymbol{\gamma})]$ is the Kullback-Leibler Divergence representing model complexity, $q(\boldsymbol{\gamma})$ is the probability distribution on $\boldsymbol{\gamma}$, is the prior distribution on $\boldsymbol{\gamma}$, E represents the data misfit

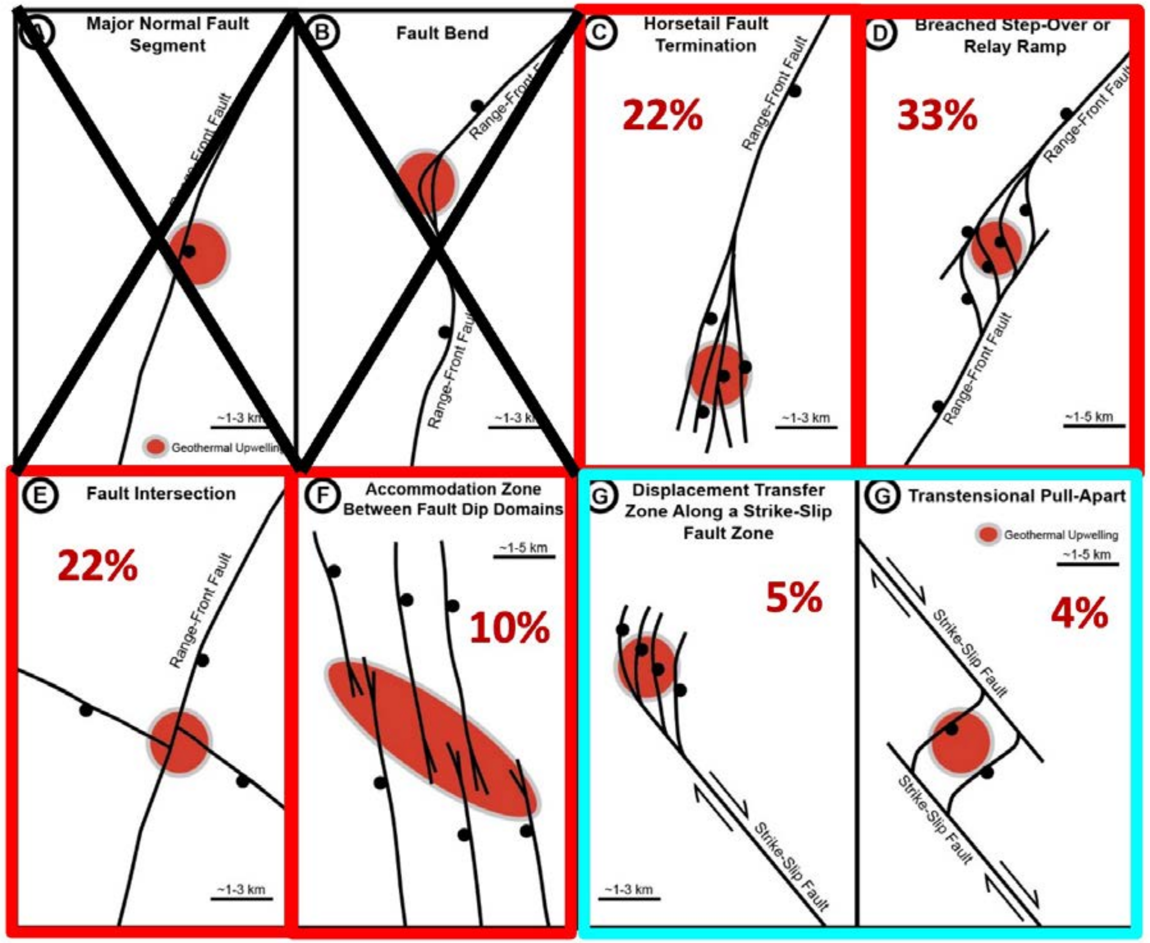
Numerical Features for Nevada Study Area*



* Some sampling bias

Categorical Features (Favorable Structural Setting)

- **450 systems analyzed; ~250 catalogued**
- **Most fields not on mid-segments of major faults**
- **Most on less conspicuous Quaternary normal faults**
- **Higher temp systems generally on faults <750 ka**
- **Hybrid settings most productive**



Faulds et al., DOE Project Review 2020