

# Model Selection II

- Philosophy of science and multiple alternative models
- Trade-offs
- Likelihood-based metrics
  - Likelihood Ratio Test
  - AIC
- Bayesian metrics
  - DIC
  - Predictive Loss

# Model selection

- Focus on choosing between multiple competing models rather than refuting a single null model
- How do we judge models?
  - Complexity
    - Number of parameters
  - Uncertainty
    - Model residuals
    - Parameter error (identifiability)
  - Data as ultimate arbiter
- “Make everything as simple as possible, but not simpler.” - A. Einstein

# Likelihood Ratio Test

- $LR = L(x|\theta_0) / L(x|\theta_1)$
- $D = -2\ln L(x|\theta_0) - -2\ln L(x|\theta_1)$
- The test statistic  $D$  is known to be distributed with a  $\chi^2$  distribution
- Degrees of freedom = Difference in # of param.
  - Overall,  $L$  increases ( $-\ln L$  declines) with # of param.
  - Penalizes model with more parameters
- $p\text{-val} = 1 - \text{pchisq}(D, df)$

# Akaike Information Criterion

$$AIC = -2 \ln L + 2p$$

- $p$  = number of parameters in the model
- Based on information theory
- Lowest value “wins”
- No p-value
- Often expressed relative to best model,  $\infty AIC$
- “Rules of thumb”
  - 0-2 = similar
  - 2-5 = weak support
  - >5 = strong

# P-value

- Probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true.
- **Not** the probability that the null hypothesis is true
  - P-value can be close to zero when the posterior probability of the null is close to 1
- **Not** the probability of falsely rejecting the null hypothesis
- **Not** biological significance

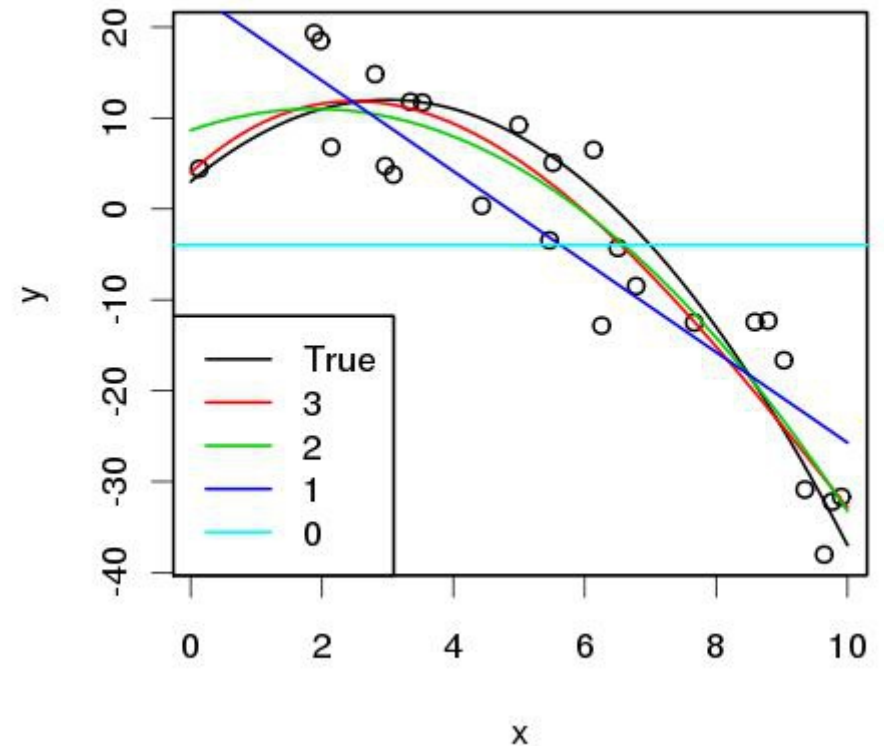
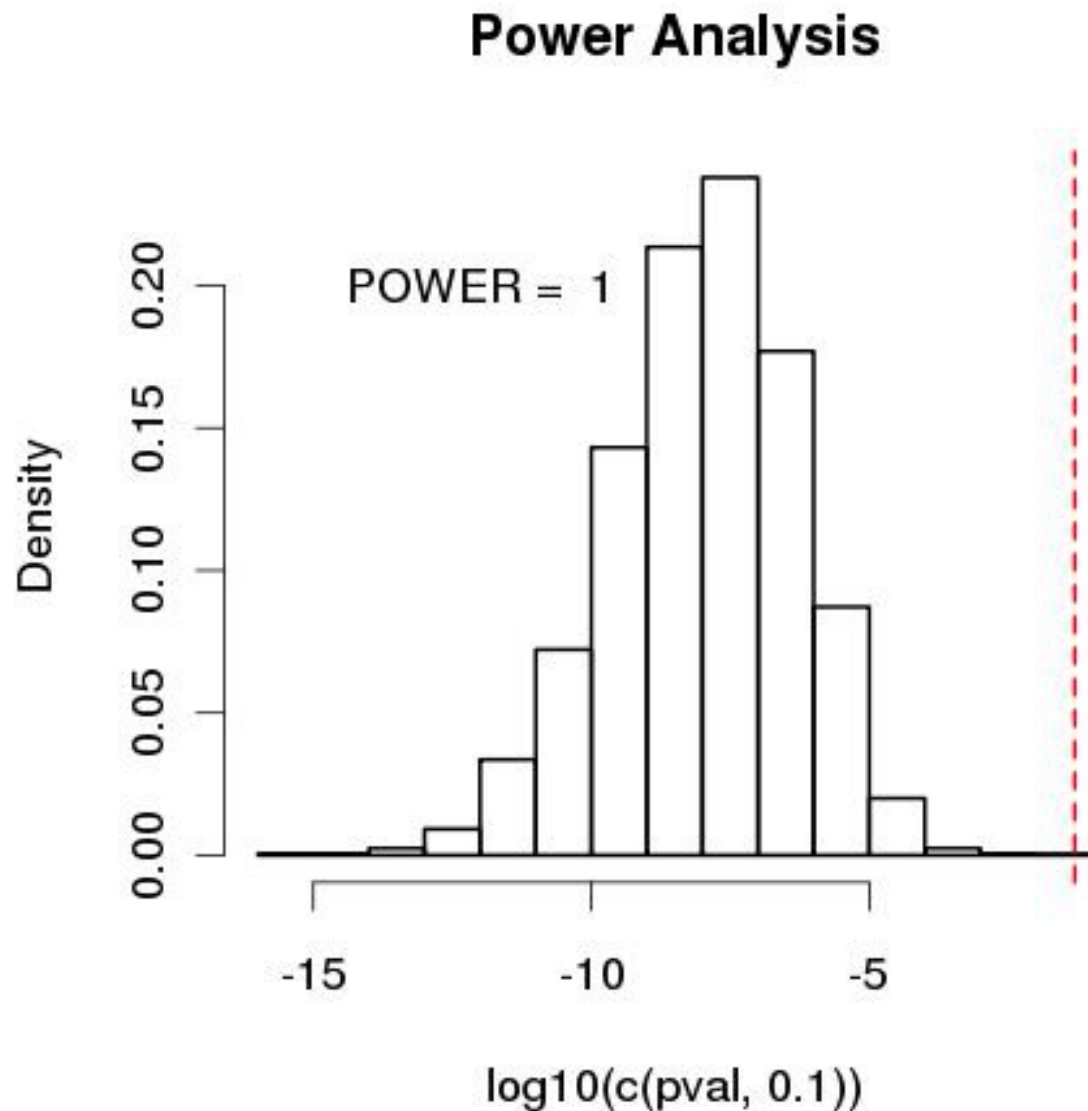
# Power

- Probability of correctly rejecting the null hypothesis
- Requires that some explicit alternative hypothesis is stated
  - Parameter values
  - Variance
  - Sample size
- Often calculated as a function of sample size
- For complex models, calculate through simulation

# Generic Example

```
LnL.A = function(theta){
  -sum(dnorm(y,f(x,theta),sd)))
}
lnL.0 = function(mu){
  -sum(dnorm(y,mu,sd))
}
for(i in 1:nsim){
  Ey = f(x,theta)      ## process model
  y = rnorm(N,Ey,sd)   ## data model
  outA = optim(ic,lnL.A) ##fit of alternative
  out0 = optim(ic,lnL.0) ##fit of null
  pval[i] = 1-pchisq(2*(outA$value-out0$value),df)
}
power = sum(pval < 0.05)/nsim
```

# Example: Quadratic vs Linear LRT



- Results **specific** to parameter values and sample size chosen



# Deviance Information Criterion

$$DIC = \bar{D} + p_D$$

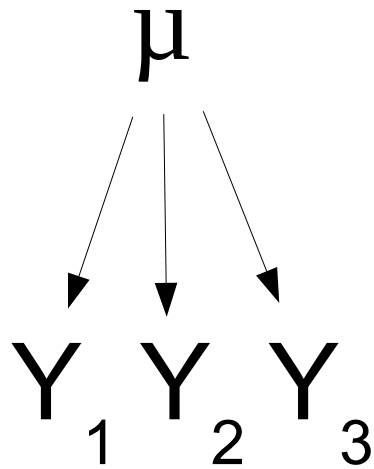
$$\bar{D} = E[D(\theta)] = \frac{1}{n_g} \sum D_i \quad p_D = \bar{D} - D(\bar{\theta})$$

- $p_D$  = effective number of parameters
- Easily calculated from MCMC
- Averages over parameter distribution rather than just single maximum
- Applicable when the number of parameters is ambiguous
- Lowest score “wins”

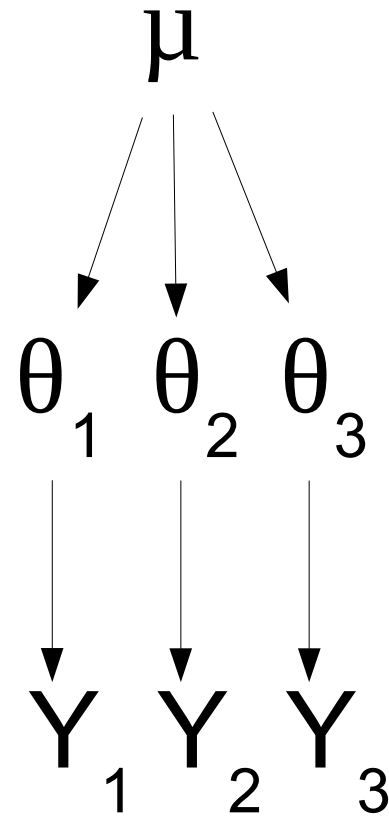
# Hierarchical Models

Hierarchical

Common mean

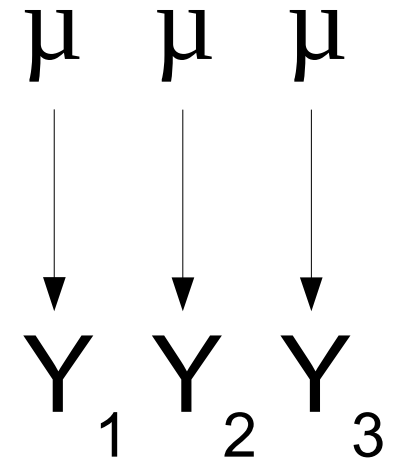


$$p=2$$



$$2 < p < 8$$

Independent



$$p=6$$

# DIC computation

```
dic.samples(model, n.iter, ...)
```

- For each MCMC iteration
  - Calculate and store deviance:  $D(\theta_i) = -2\ln L(y|\theta_i)$
- After MCMC
  - Calculate posterior means for parameters  $\bar{\theta}$
  - Calculate D at  $\bar{\theta}$  :  $D(\bar{\theta}) = -2\ln L(y|\bar{\theta})$
  - Calculate  $\overline{D(\theta)} = \sum D(\theta_i) / n_g$
- $DIC = 2\overline{D(\theta)} - D(\bar{\theta})$

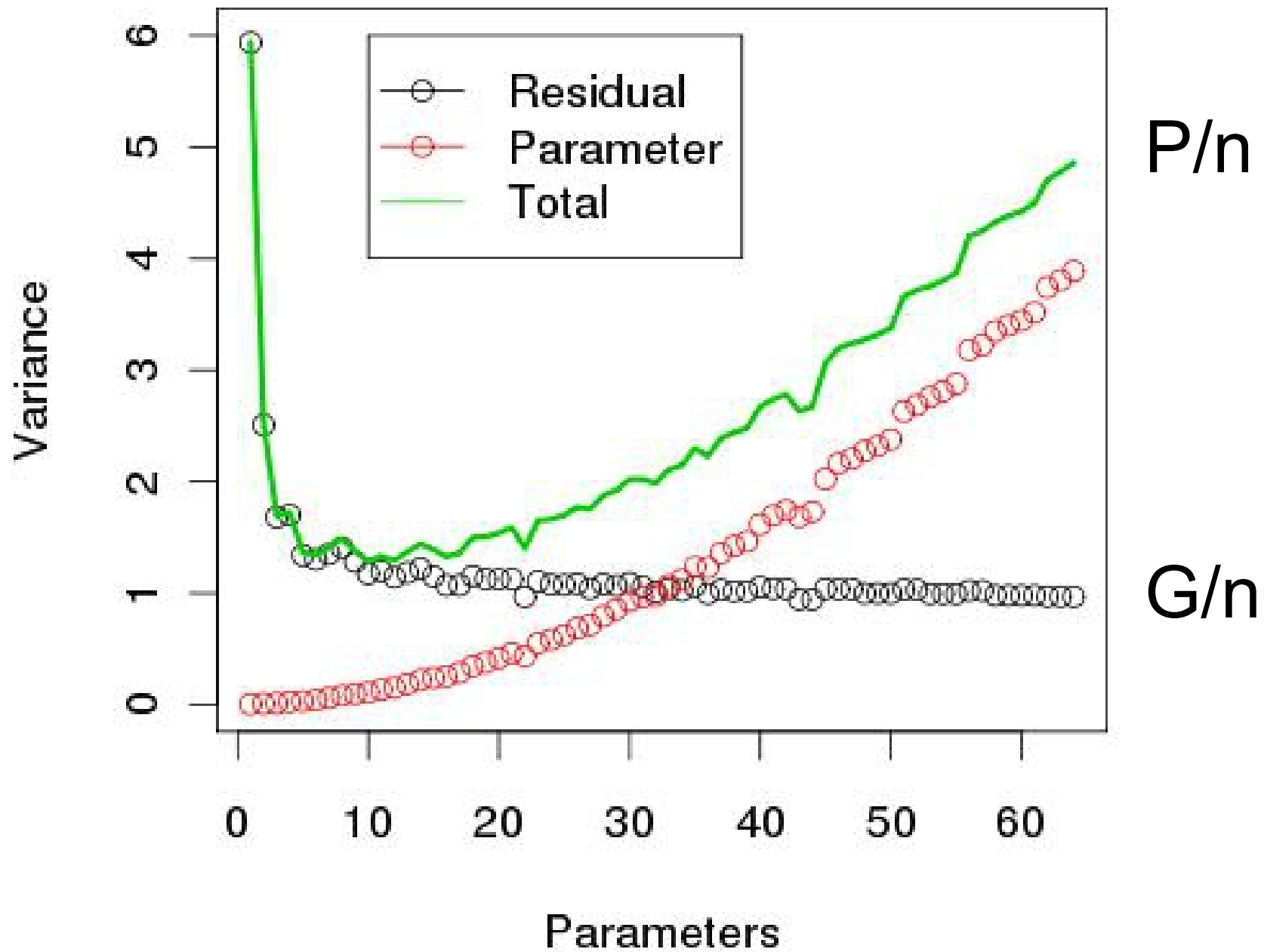
Model	DIC	pD	$\Delta$ DIC
flat	221.10	2.06	82.00
linear	174.40	3.12	35.30
quadratic	139.10	4.15	0.00
cubic	141.40	5.27	2.30

# Predictive Loss

$$D_{pl} = G + P$$

- $G$  = total residual SS  $\sum (E[y_{rep}] - y_{obs})^2$
- $P$  = total predictive variance  $\sum var[y_{rep}]$
- Given  $y_{obs}$ , predict replicate  $y_{rep}$ 
  - i.e. predictions made for same points as observations
- Focused on prediction, easily calc from MCMC
- Does not require model dimension

# UNCERTAINTY



# Predictive Loss Algorithm

- For every MCMC step
  - Generate pseudodata at same points/covariates as the original data (otherwise equiv. PI calc.)
- From posterior predictive distribution
  - Calculate posterior mean for each point:  $E[y_{rep}]$
  - Calculate residual variance for each point:  $Var[y_{rep}]$
  - $P = \text{sum of } Var[y_{rep}] \text{ over all points}$
  - $G = \sum (E[y_{rep}] - y_{obs})^2$

# Predictive Loss: Quadratic

model	P	G	D
flat	10065.16	8596.29	18661.45
linear	1546.03	1215.26	2761.3
quadratic	378.68	272.7	651.38
cubic	410.45	271.3	681.74

Note:  $\sqrt{P/n}$  and  $\sqrt{G/n}$  are the predictive SD and residual SD respectively



# Bayes Factor

$$BF = \frac{p(M_1|y)/p(M_2|y)}{p(M_1)/p(M_2)}$$

- Require assigning a prior probability to each model
- Hard to calculate except in limited cases
- Asymptotically tends to select too simple
- I have not seen BF used much recently

# Bayesian Model Averaging

- Make predictions using all of your alternative models

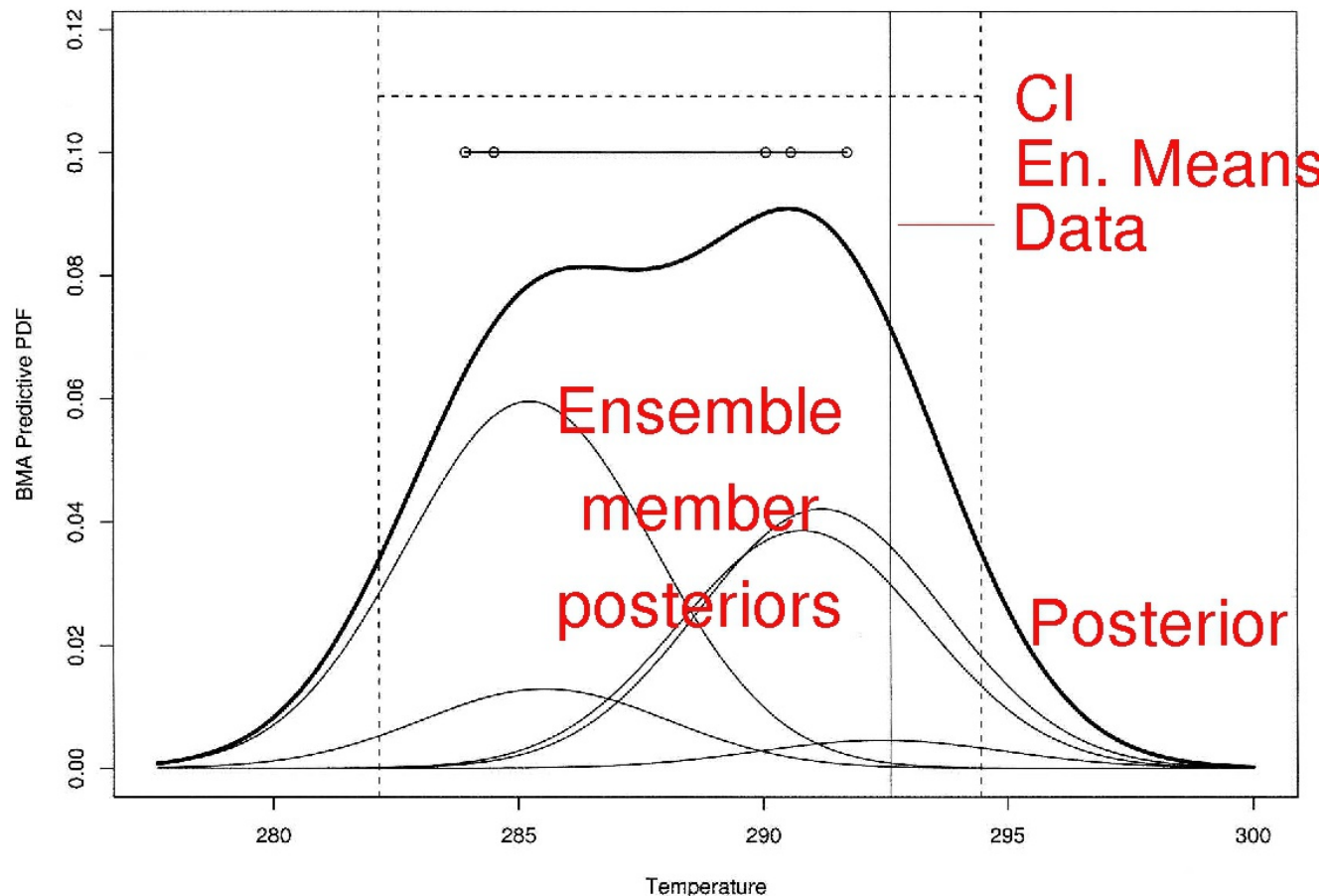


FIG. 3. BMA predictive PDF (thick curve) and its five components (thin curves) for the 48-h surface temperature forecast at Packwood, WA, initialized at 0000 UTC on 12 Jun 2000. Also shown are the ensemble member forecasts and range (solid horizontal line and bullets), the BMA 90% prediction interval (dotted lines), and the verifying observation (solid vertical line).

# Reversible Jump MCMC

- Considers the number of terms in a nested model to be unknown
- Will add and remove terms within the MCMC step
- Generates a posterior probability for each model
- Prediction automatically averages over models
- “in fashion”

# As of now we can...

- Fit models (Likelihood and Bayes)
- Construct Confidence intervals
- Test Hypotheses / compare models
- Make predictions that propagate uncertainty

## What's left?

- Exploration of common and more advance models
  - Useful approaches/models for certain types of problems