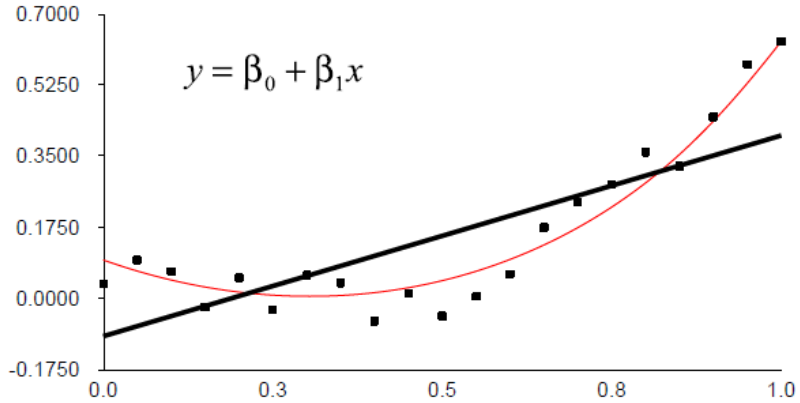
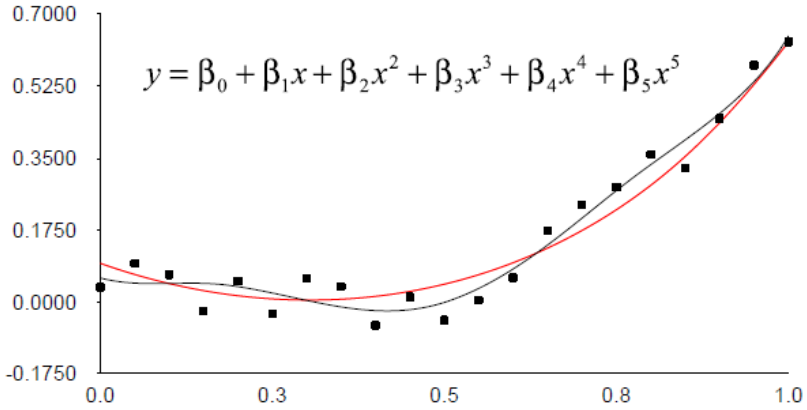
Model assessment

Jason McLachlan, Notre Dame

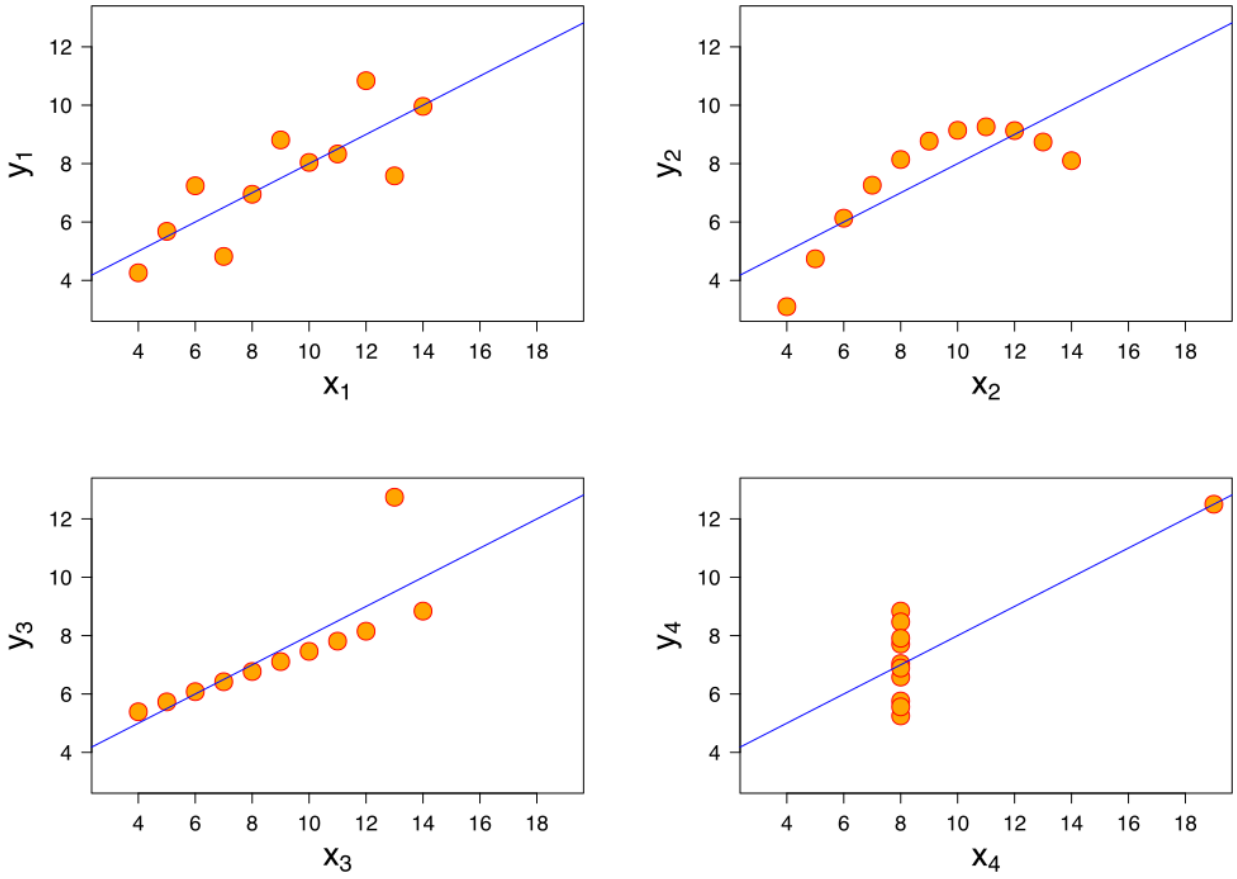


Two few parameters--fails to respond to information. Bias is high.



Too many parameters--responds to "noise." Variance is high.

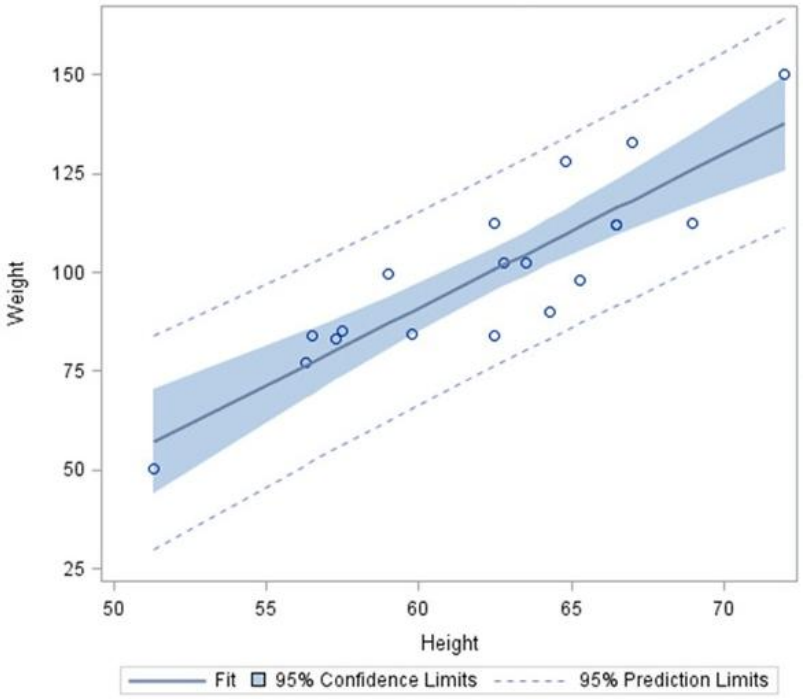
Let's start in familiar territory: regression model



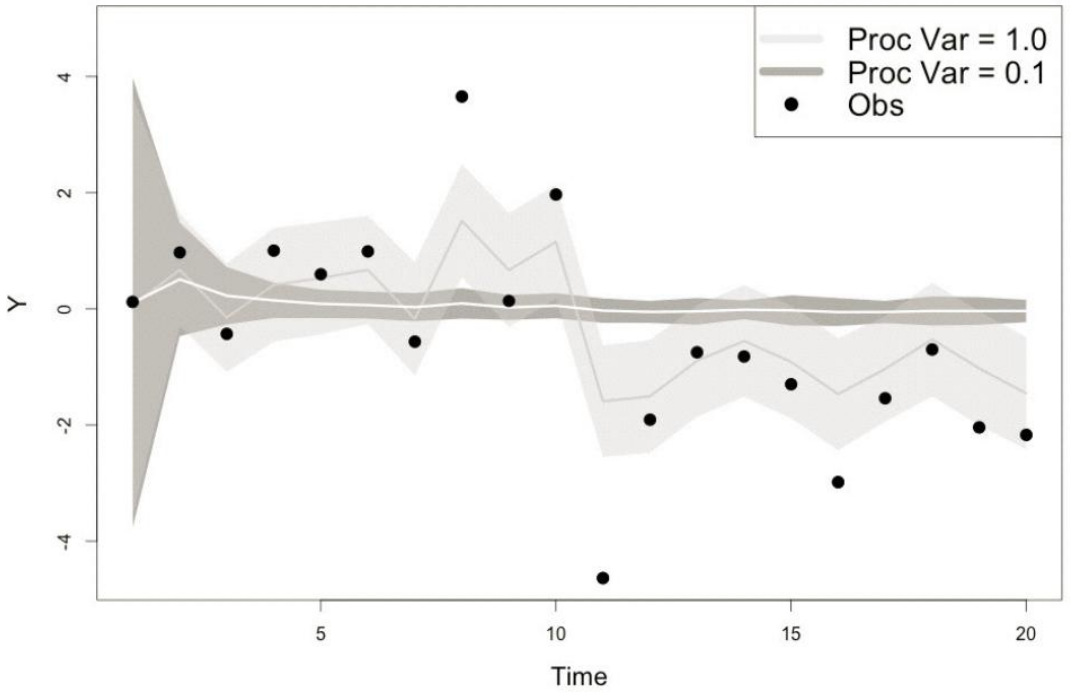
Why do we assess model fit?
How do we assess model fit?
What do we learn from this?

How much of this intuition can we take with us?

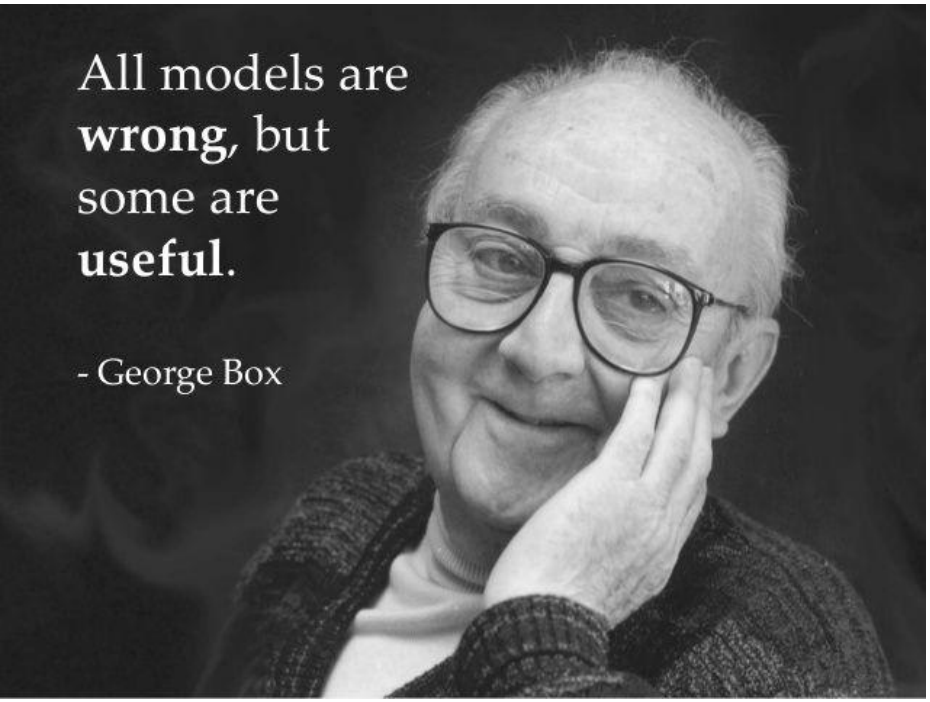
LINEAR REGRESSION



FORECAST (e.g. filter divergence)



What's the difference?
What has changed?
What has stayed the same?



Hierarchical Bayes and forecasting can seem very open, and it can be hard to see what your best choice is.

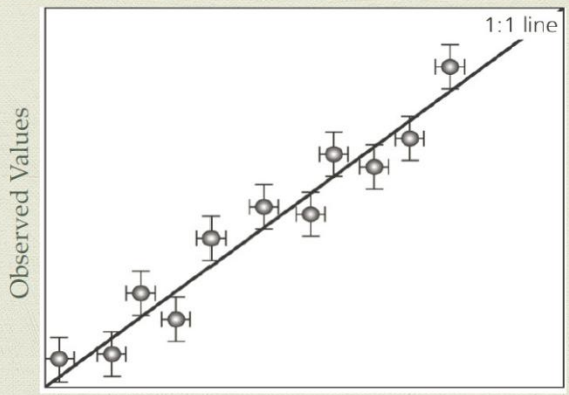
Modeling is a free and creative process.

Don't worry about getting it right (it's not), but how can you make sure that your model is useful?

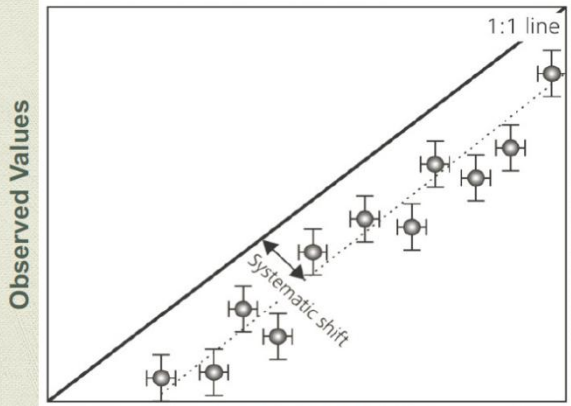
(A) Can it give rise to new observations that properly resemble the original data?
(B) Does it add to our knowledge, resolve disputes, or add depth of understanding?
(C) Does it make useful forecasts?

Can the model give rise to new observations like the data?

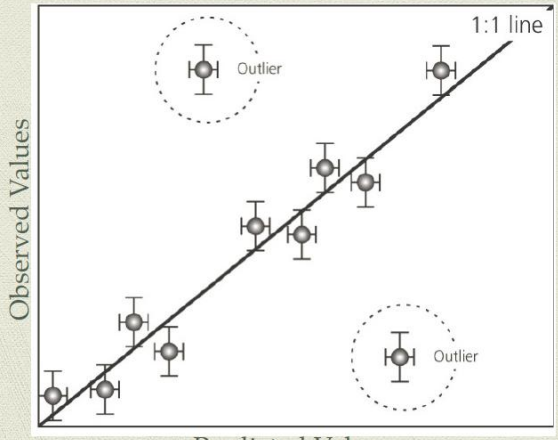
Accuracy of Prediction



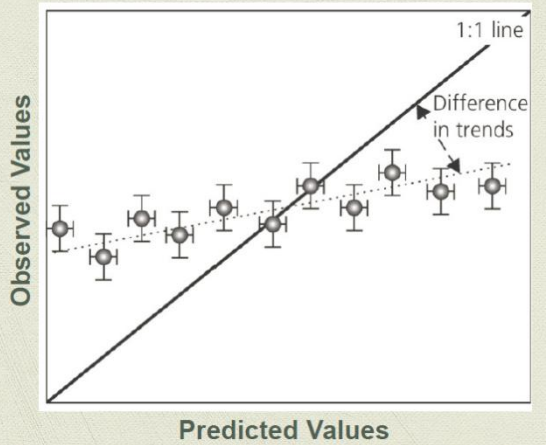
Assess Biases



Identify Outliers

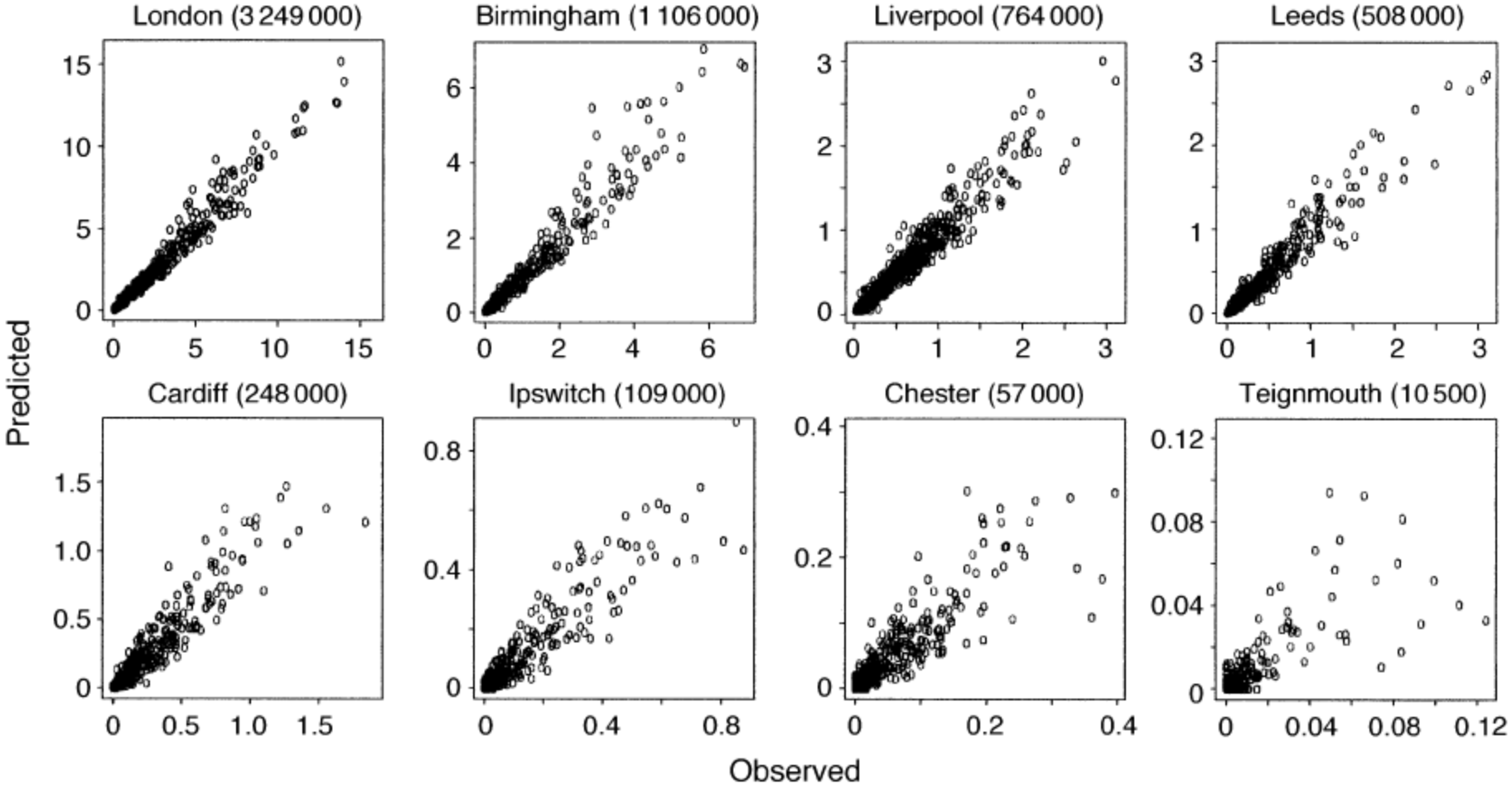


Miscalibration



It's easy to focus on getting a good fit.
But you learn more from the surprises.

Can the model give rise to new observations like the data?



Can the model give rise to new observations like the data?

We can do this more formally thanks to a key feature of Bayesian inference: everything (except the data) is a random variable. We can sample from the posterior:
- **Predictions, including missing data (forecasts)**
- **Latent variables**
- **Derived quantities**
- **etc**

Check residuals

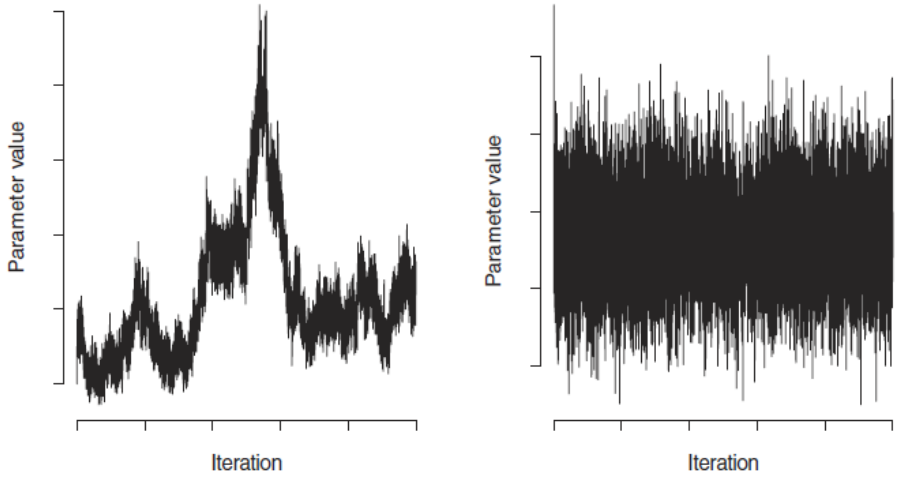
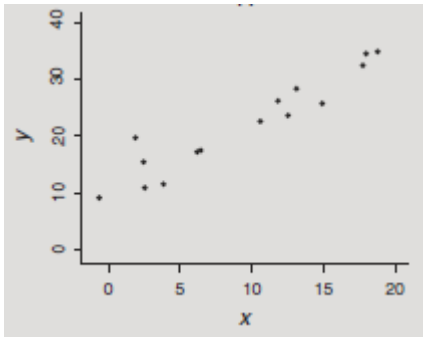


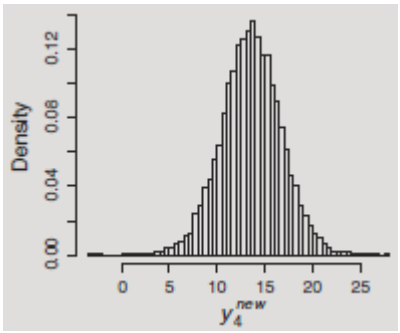
Figure 7.3.4. A trace plot of a chain that has not converged will "wander" among different values of the parameter, so that the band of values is not level and well mixed (left panel). Compare with the trace plot of the converged chain in the right panel.

Predict new data (y^{new})from model via the posterior predictive distribution

$$[\boldsymbol{\theta}, y^{new}|\mathbf{y}] \propto [\mathbf{y}|\boldsymbol{\theta}, y^{new}]\,[\boldsymbol{\theta}, y^{new}]$$

$$[y^{new}|\mathbf{y}] = \int [y^{new}|\boldsymbol{\theta}]\,[\boldsymbol{\theta}|\mathbf{y}]\,\mathrm{d}\boldsymbol{\theta}$$





(Hobbs & Hooten 2015)

Can the model give rise to new observations like the data?

Bayesian p-values (P_B)
Choose the things you care about.
Make posterior predictive draws.
Compare their distribution to the data.

Conventional p-value:
Probability of observing a more extreme test statistic than that calculated from the data.

Posterior predictive check:
Probability that distribution of data generated by our model is more extreme than the distribution of the data.

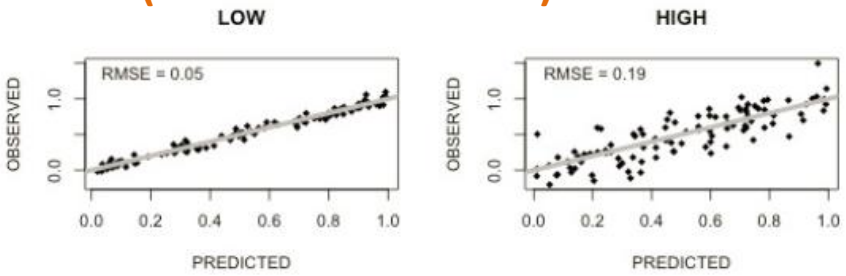
$$P_B = \Pr\left(T(\mathbf{y}^{new}, \boldsymbol{\theta}) \geq T(\mathbf{y}, \boldsymbol{\theta})|\mathbf{y}\right)$$

What test statistic, T(y, θ), should we use?
Whatever we care about: mean, variance, kurtosis, maximum value, chi-squared value, …

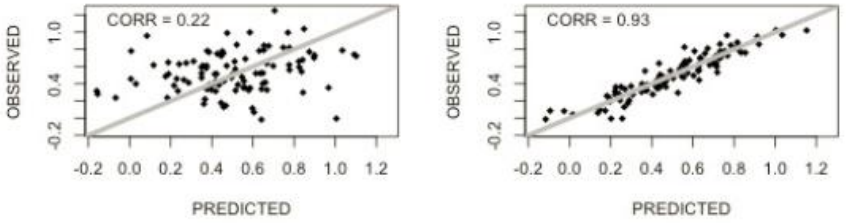
This is easy to do. Just sample y^{new} at each MCMC iteration, and compare to data.

Can the model give rise to new observations like the data?

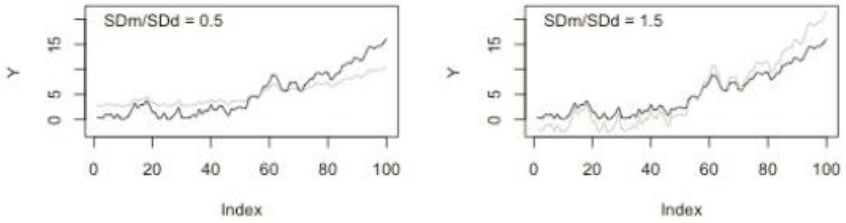
RMSE (std dev of residuals)



Correlation coefficient



SD_{model}/SD_{data} (under or overpredict?)



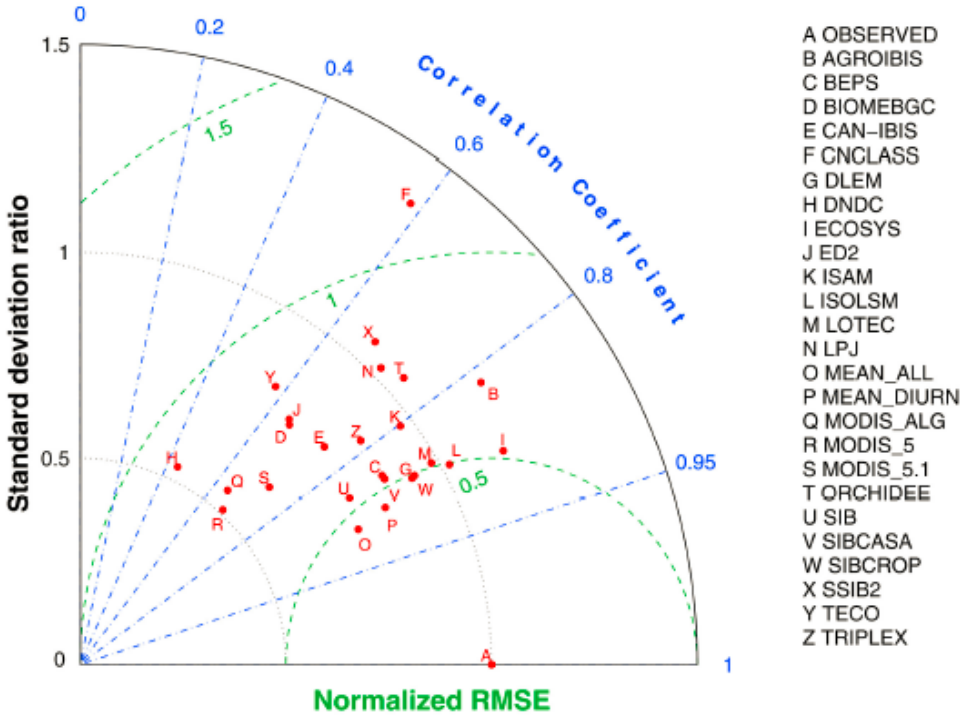
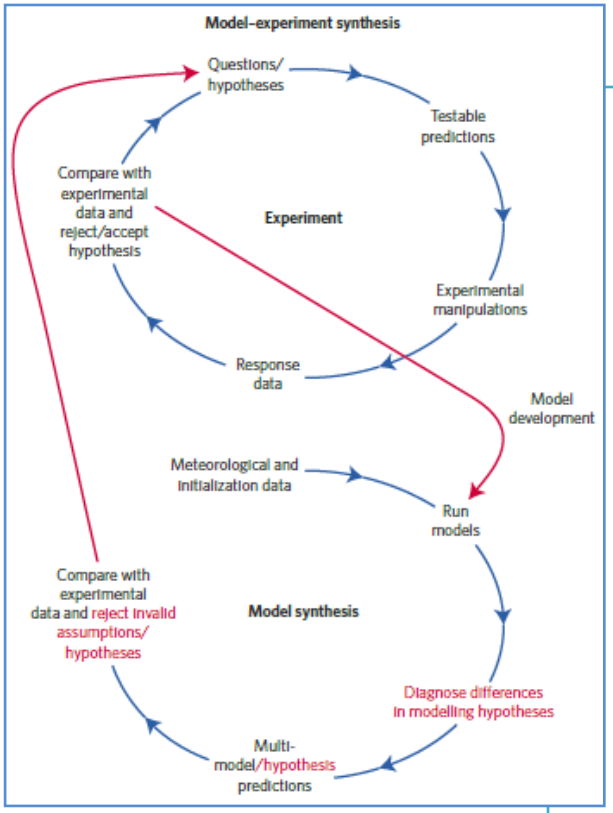
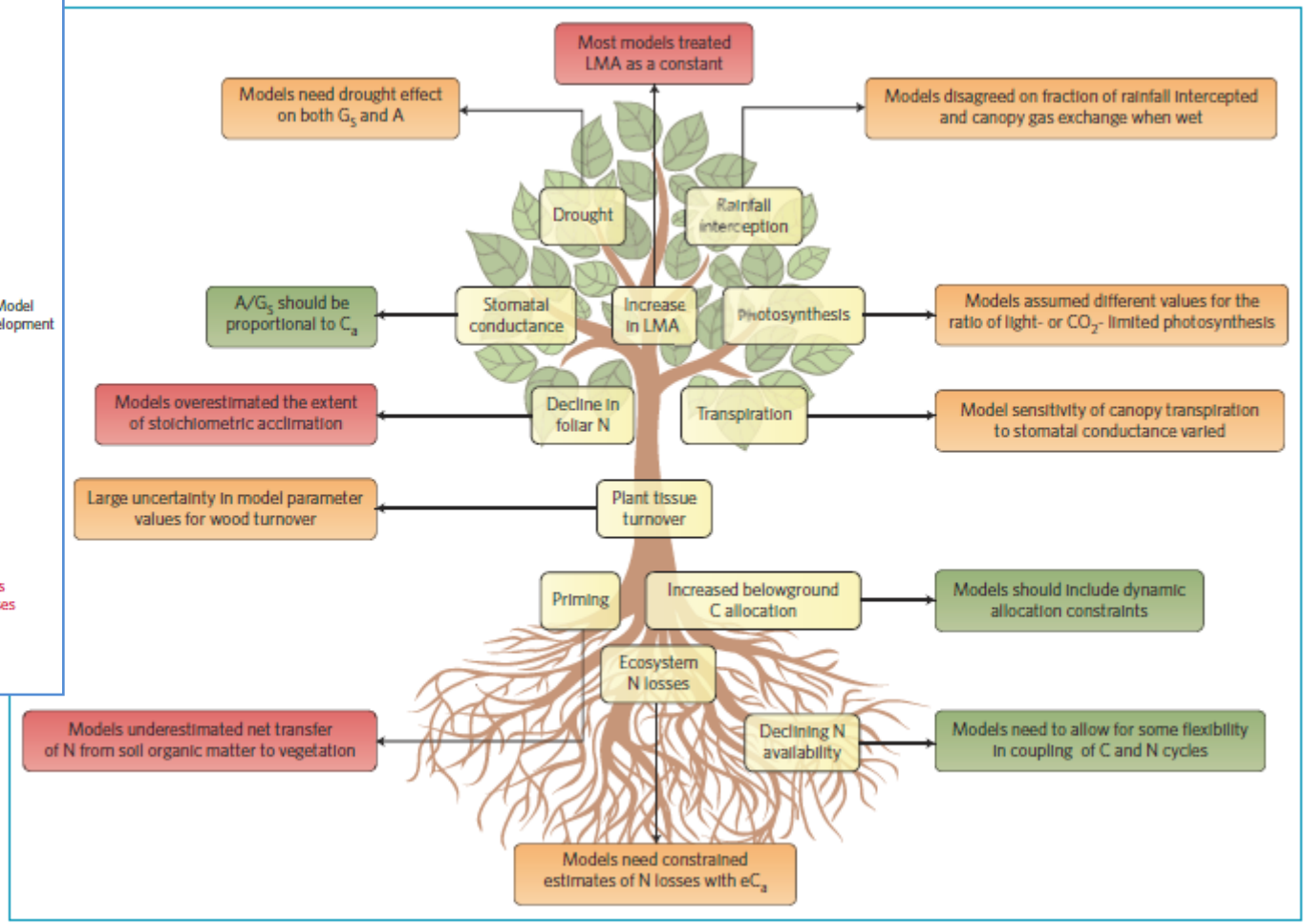


Figure 16-3 Taylor Diagram assessing the ability of 20 ecosystem models to predict daily mean GPP across 39 eddy-covariance towers. Data are located at point A. The multi-model ensemble mean (O, P) had the lowest RMSE and highest correlation but underpredicted the true variability. A cluster of models (G, M, L, W) had a slightly lower correlation and higher RMSE but had similar variability to the observations. The MODIS GPP products (Q, R, S) performed worse than most models and substantially underpredicted the true variability in the system. (Schaefer *et al.* 2012)

Does the model add depth of understanding?





Data allowed discrimination of alternative assumptions

More data needed to discriminatie alternatives

Data identified missing or wrong assumptions

(Medlyn et al. 2015)

Does the model add depth of understanding?

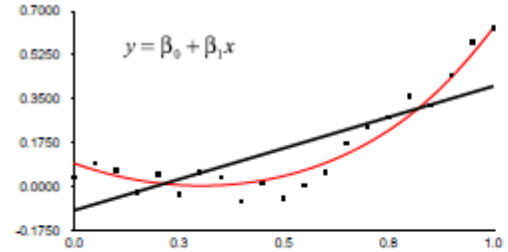
Inference from Multiple Models

A guide to Bayesian model selection for ecologists

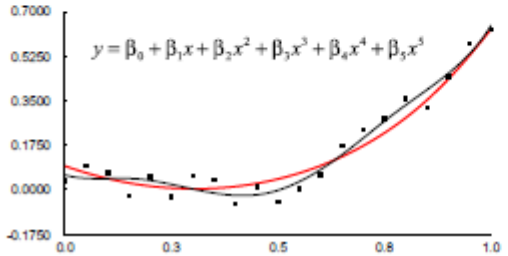
M. B. Hooten[1,2,3,4,7] and N. T. Hobbs[4,5,6]

Information criteria regularize an optimization, balancing model fit, with model parsimony

$$\underbrace{\mathscr{L}(\mathbf{y}, \boldsymbol{\theta})}_{\text{loss function}} + \underbrace{r(\boldsymbol{\theta}, \boldsymbol{\gamma})}_{\text{regulator}}$$



Two few parameters--fails to respond to information. Bias is high.



Too many parameters--responds to "noise." Variance is high.

Example Watanabe-Akiake Information Criterion (WAIC)

Pointwise predictive score
gets smaller with more parameters

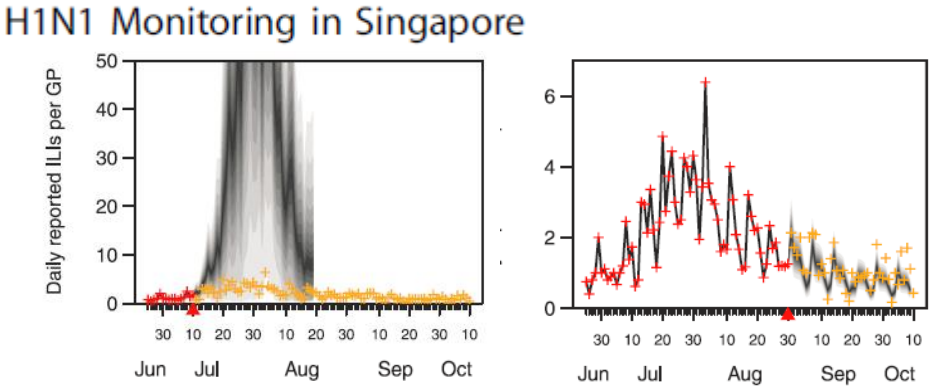
$$\text{WAIC} = \overbrace{-2\sum_{i=1}^{n} \log \underbrace{\int [y_i|\boldsymbol{\theta}][\boldsymbol{\theta}|\mathbf{y}] d\boldsymbol{\theta}}_{\text{Hey! It's our old friend the posterior predictive distribution.}}}^{\text{gets smaller with more parameters}} + \underbrace{\sum_{i=1}^{n} \text{Var}_{\boldsymbol{\theta}|\mathbf{y}}(\log[y_i|\boldsymbol{\theta}])}_{\text{gets larger with more parameters}}$$

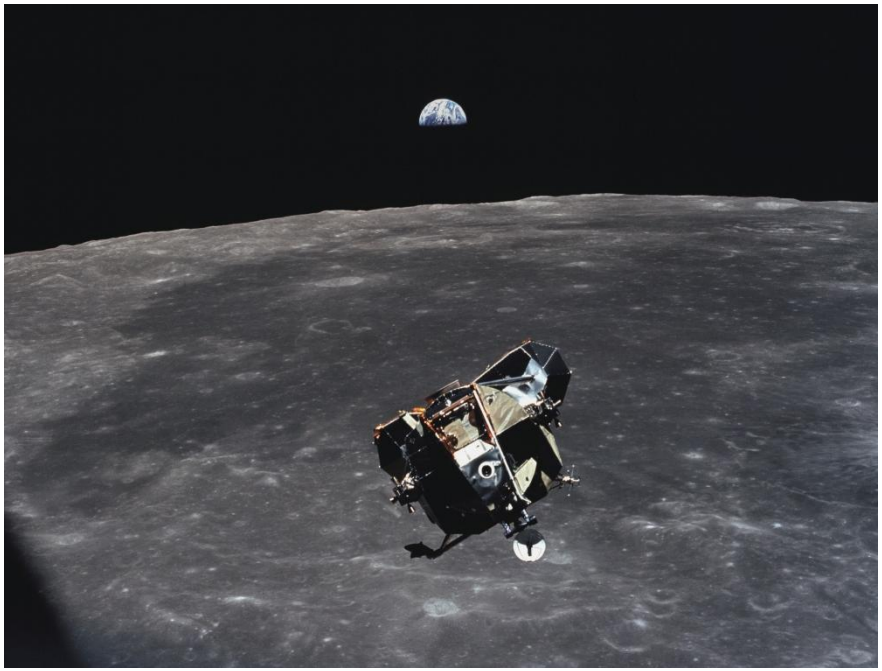
Effective number of parameters

Does the model add depth of understanding?

By far, the best way to assess or select among models is to predict out of sample (oos)!
For example, using mean squared prediction error:

$$\text{MSPE} = \sum_{i=1}^{n} \frac{(y_{\text{oos},i} - \hat{y}_{\text{oos},i})^2}{n}$$





“Data assimilation is not rocket science, but it can be used for that.”

- Dave Moore

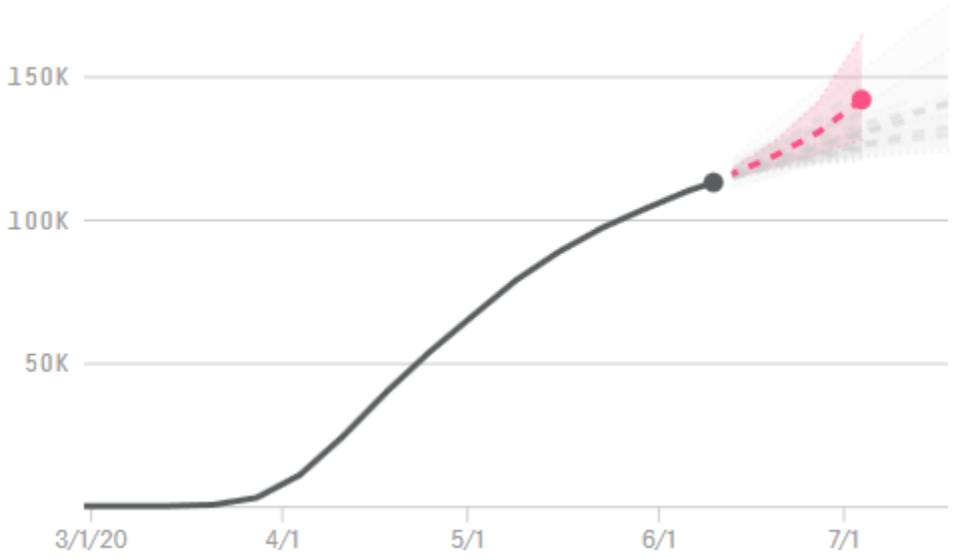
“History is a sequence of random events and unpredictable choices which is why the future is so difficult to foresee. But you can try.”

- Neil Armstrong

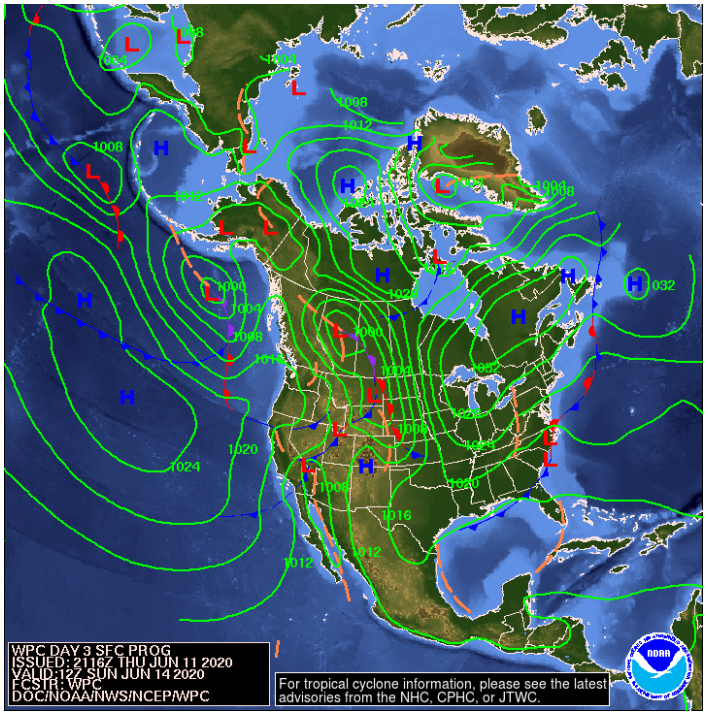
Does the model make a useful forecast?

A STRATEGIC VISION FOR NOAA'S ECOLOGICAL FORECASTING ROADMAP

2015-2019



USA npn National Phenology Network



Thanks!