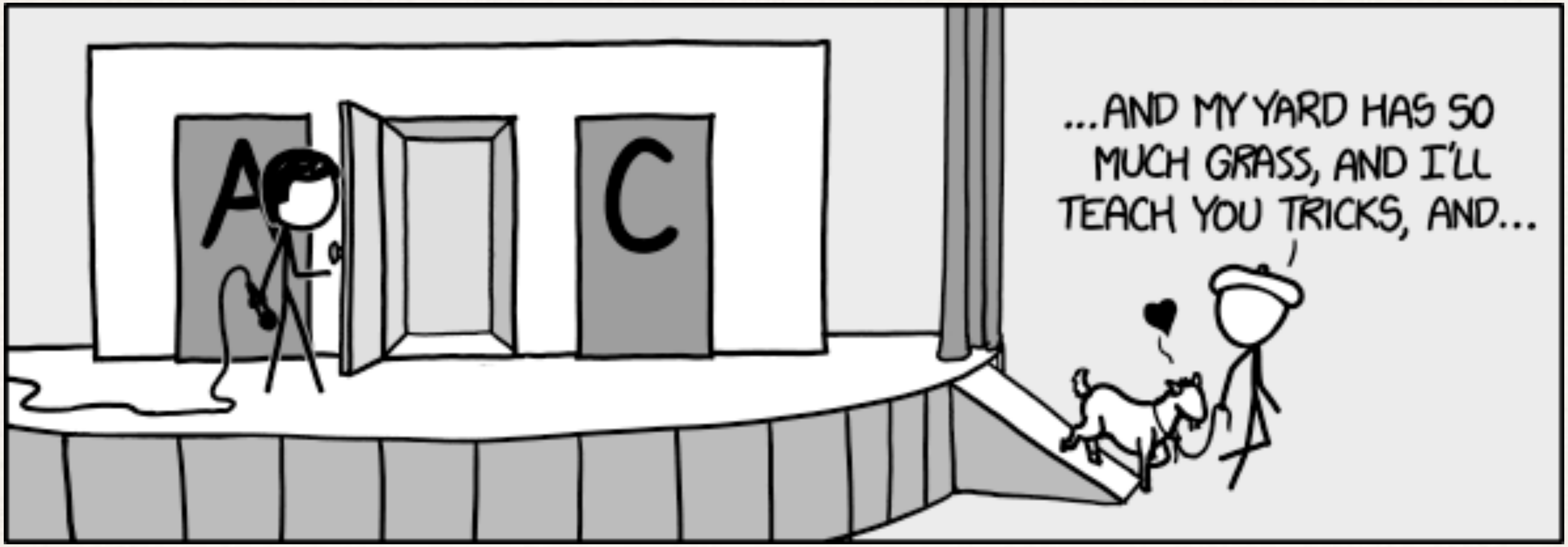
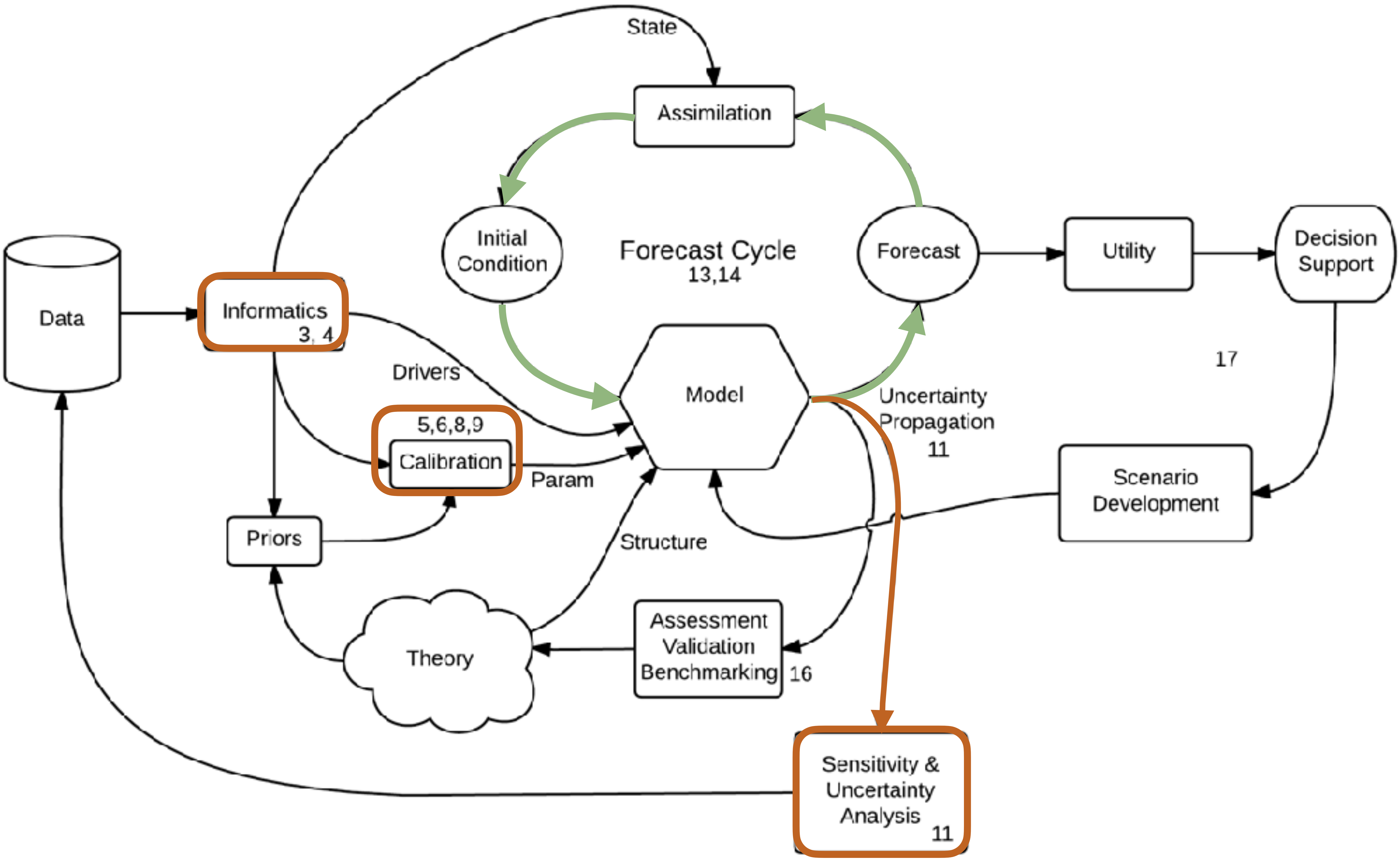


Data Assimilation 1: Analytical Methods



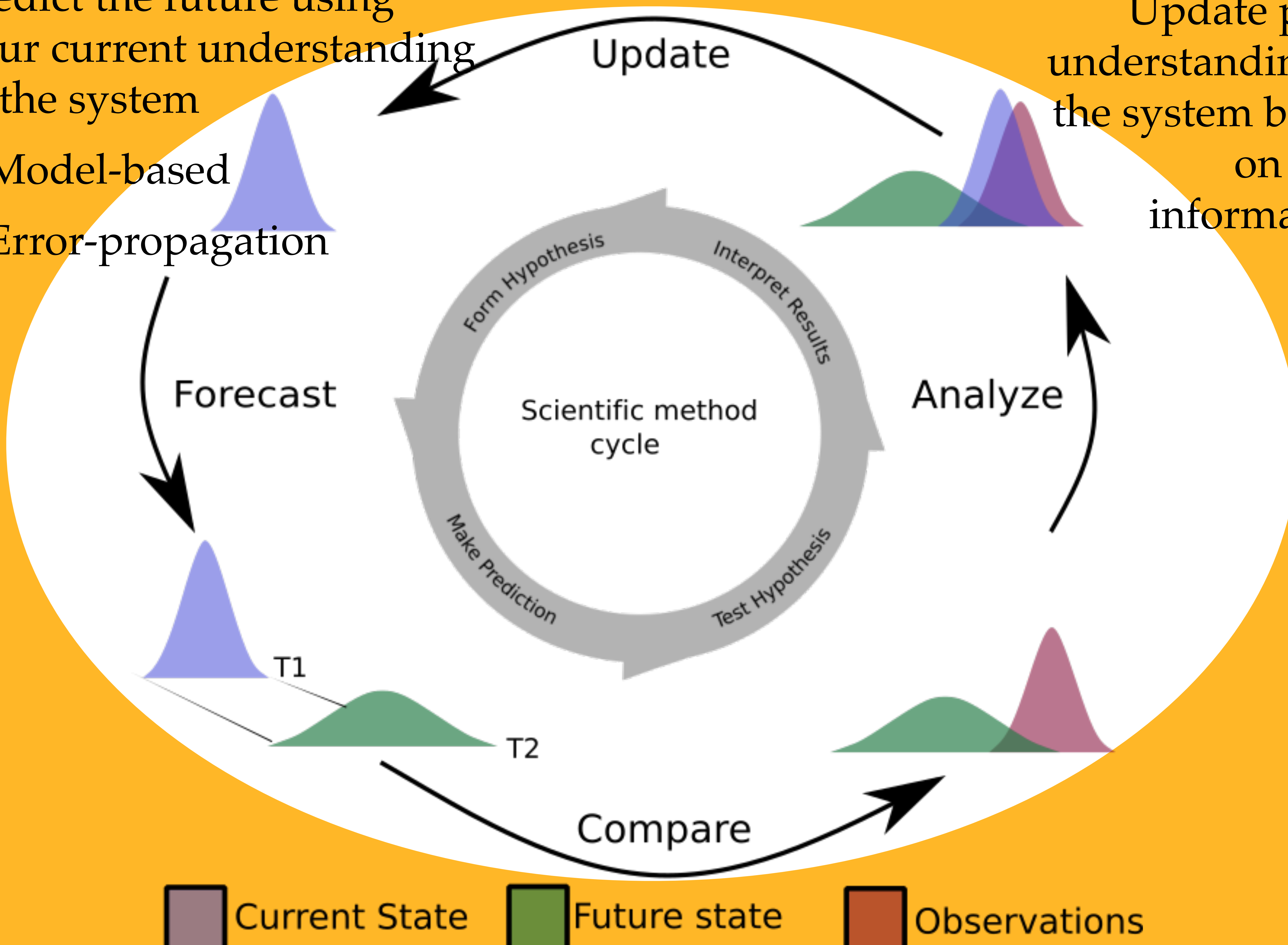
State
Assimilation
Initial Condition
Forecast Cycle
13,14
Forecast
Utility
Decision Support
Data
Informatics
3, 4
Drivers
Model
17
Uncertainty Propagation
11
5,6,8,9
Calibration
Param
Scenario Development
Priors
Structure
Assessment Validation Benchmarking
16
Theory
Sensitivity & Uncertainty Analysis
11

FORECAST-ANALYSIS CYCLE

Predict the future using your current understanding of the system

- Model-based
- Error-propagation

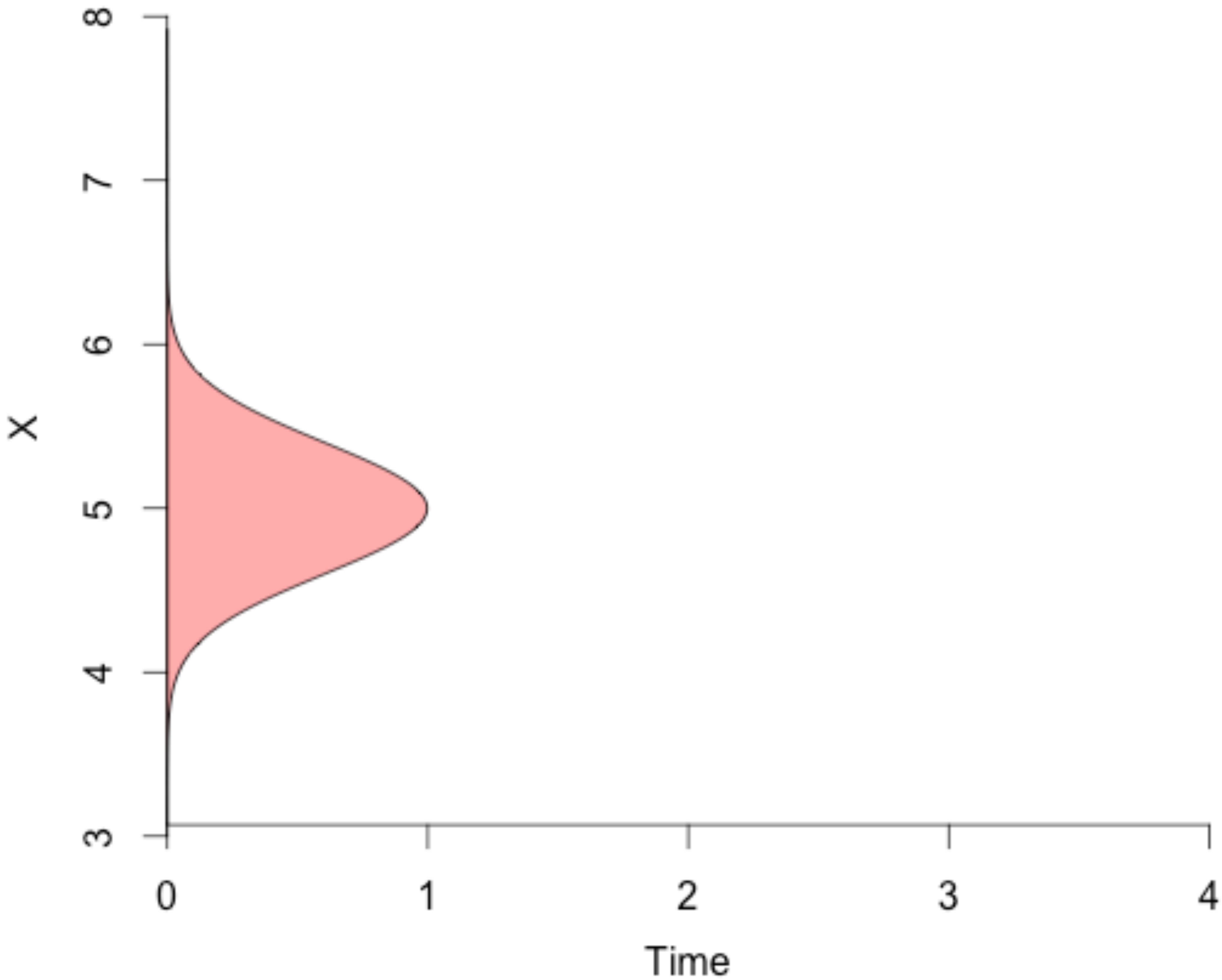
Update prior understanding of the system based on new information

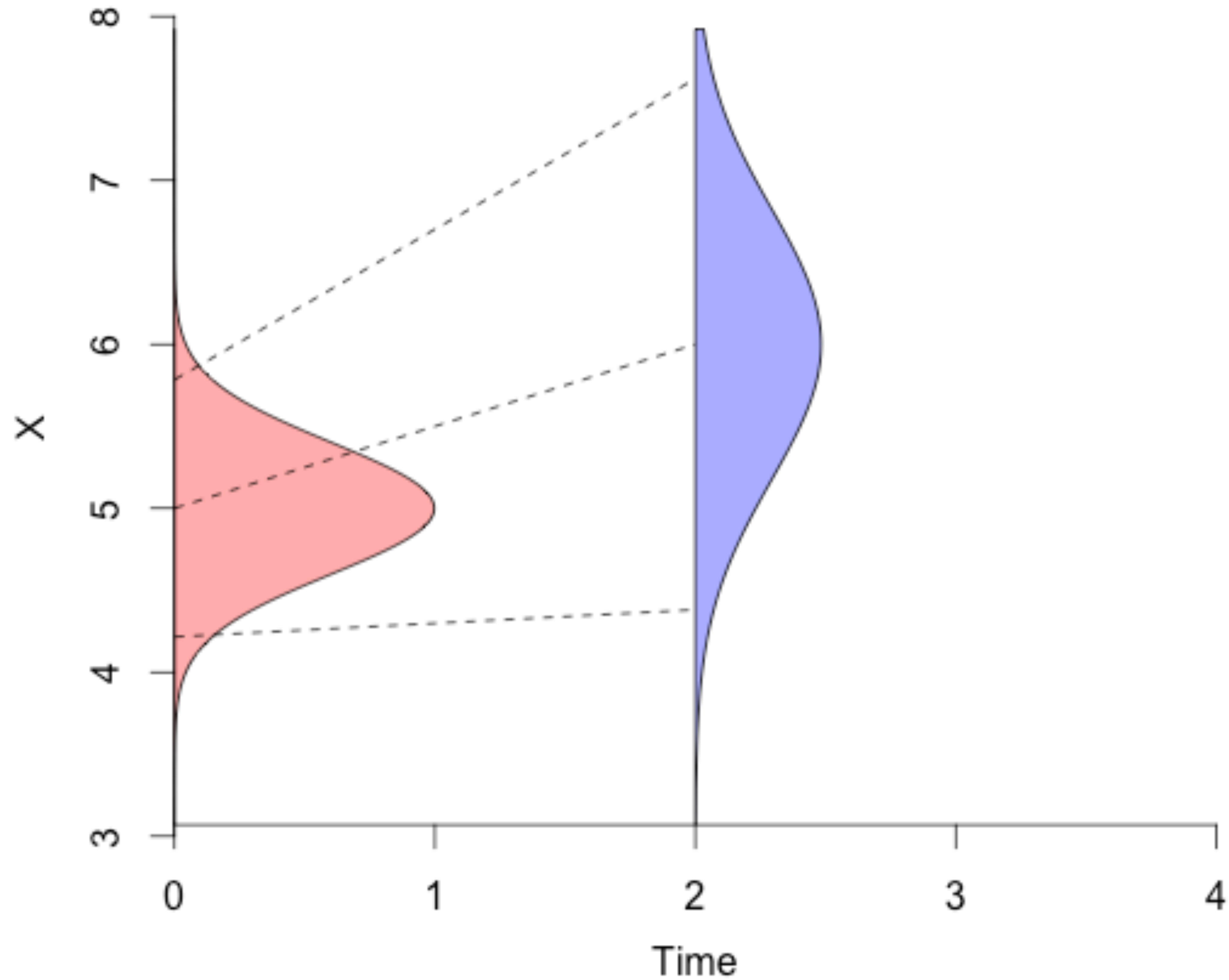


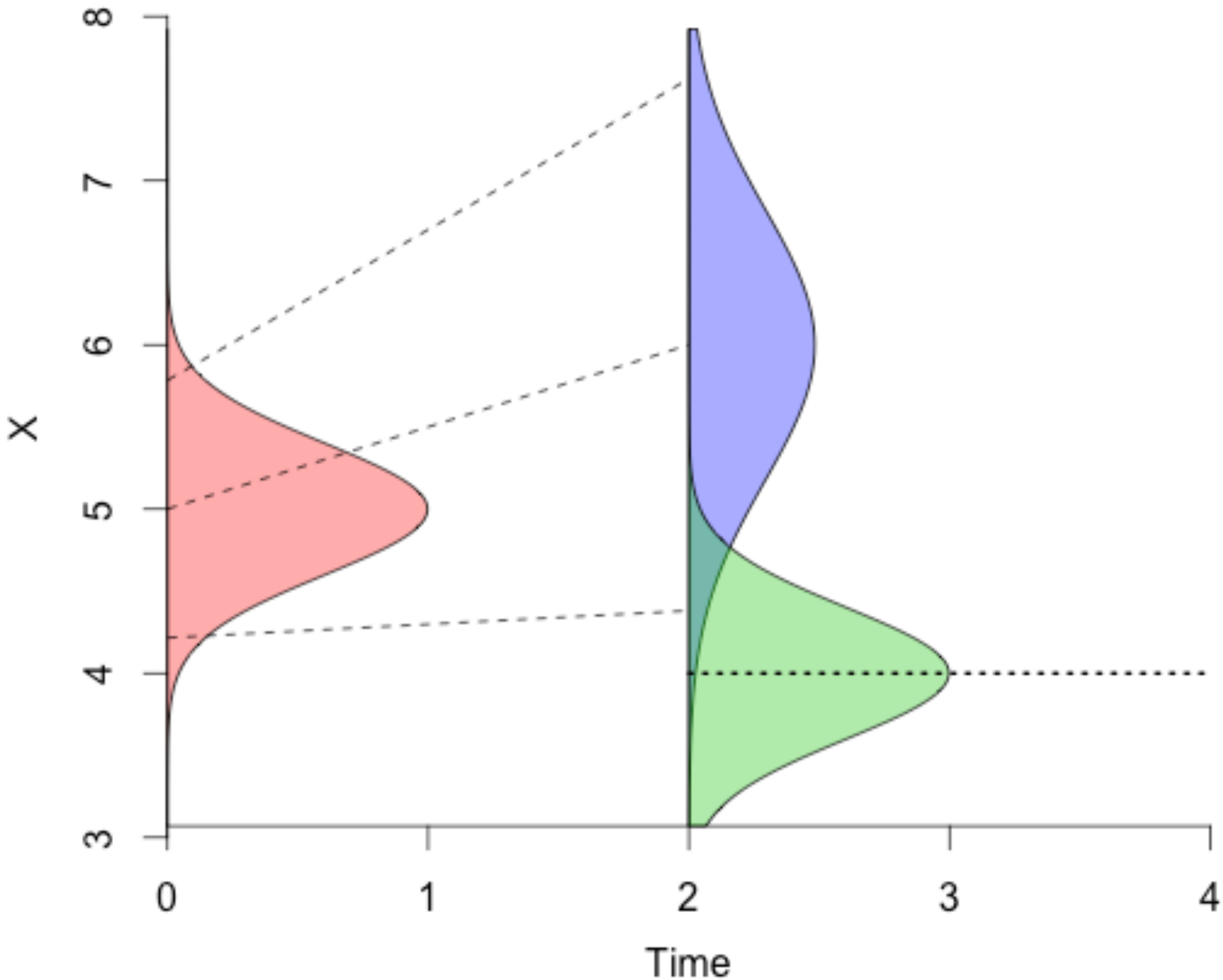
The Analysis Problem

- Prior to observing how the future plays out, what is our best estimate of the future state of the system, X_{t+1}?
  - The forecast, $P(X_{t+1})$
- Once we make (imperfect) observations of the system, Y_t, what's our best estimate of X_t?
  - $P(X_{t+1}) = P(Y_{t+1})$?
  - $P(X_{t+1} \mid Y_{t+1}) \propto P(Y_{t+1} \mid X_{t+1})\, P(X_{t+1})$

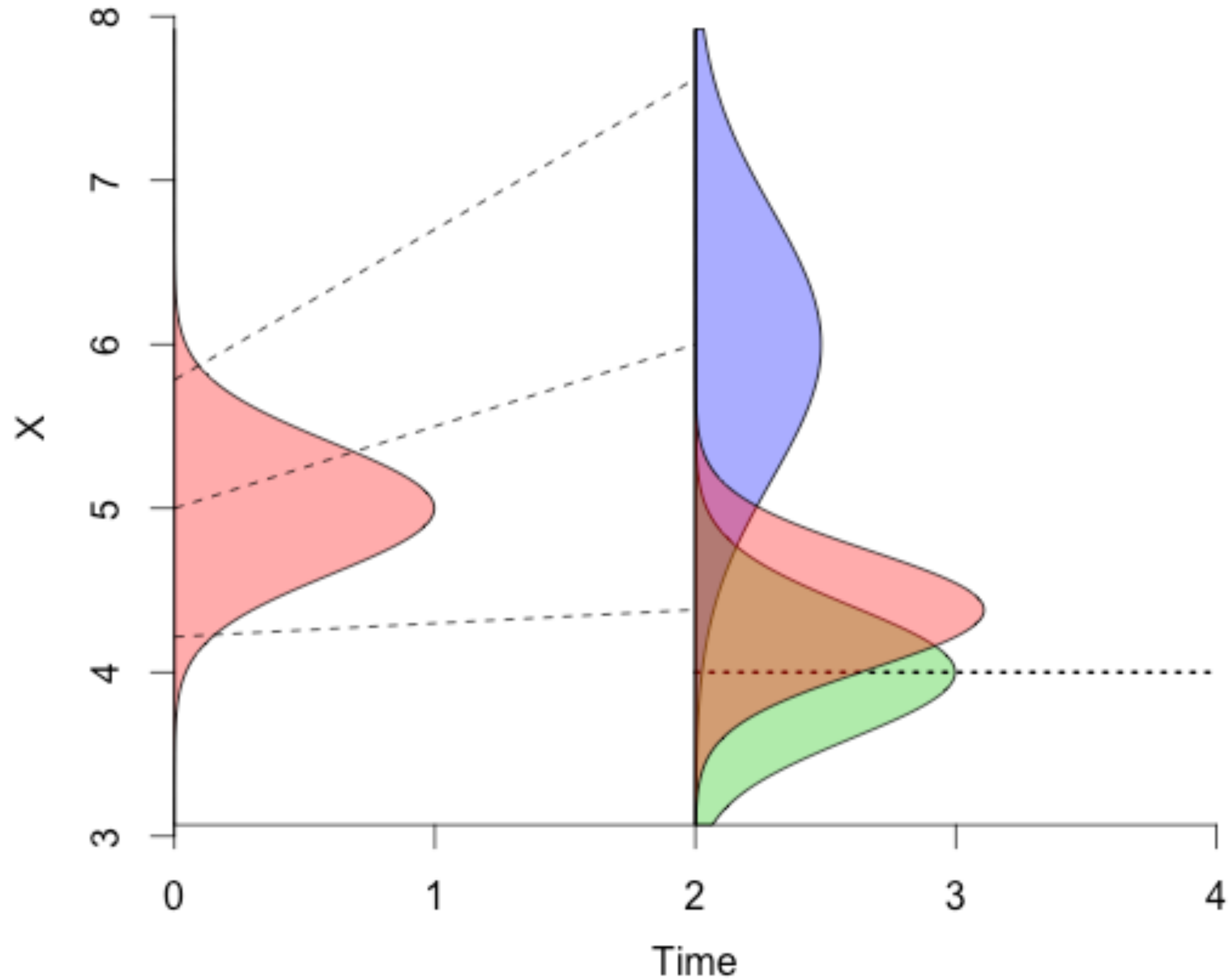
Posterior Likelihood Prior


8
7
6
5
4
3
X
0
1
2
3
4
Time

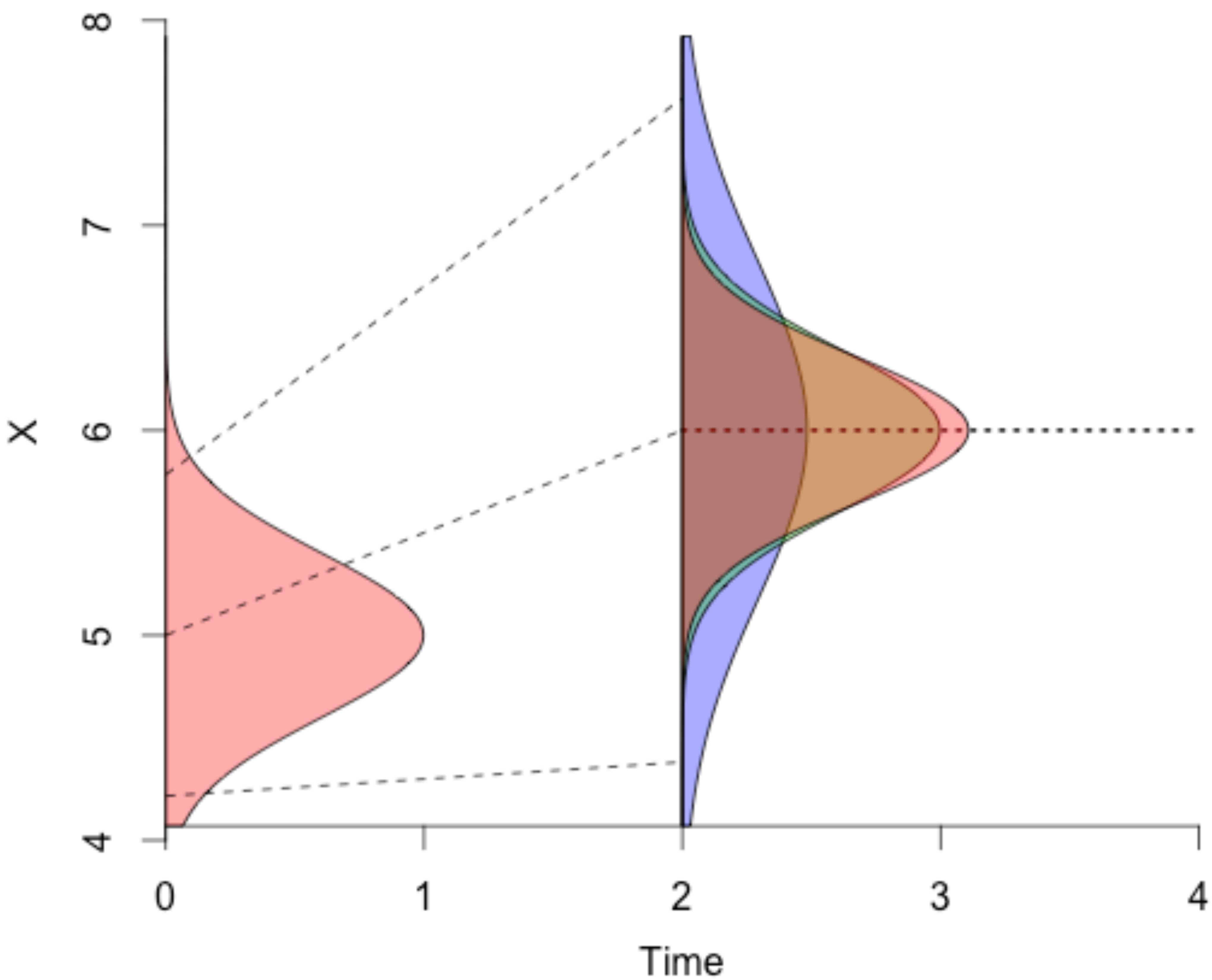

X
8
7
6
5
4
3
0
1
2
3
4
Time



8
7
6
5
4
3
X
0
1
2
3
4
Time

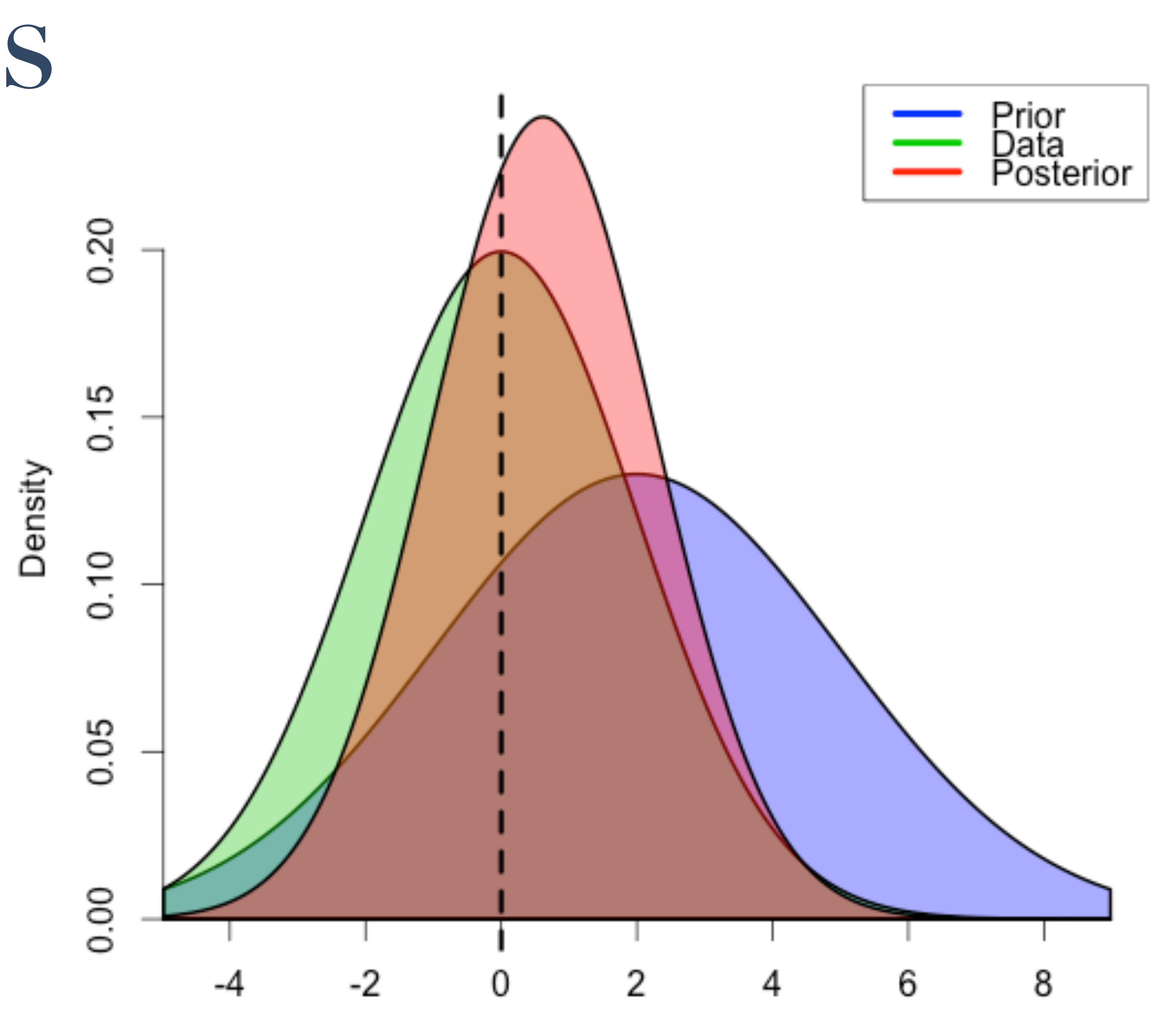


8
7
6
5
4
3
X
0
1
2
3
4
Time



X
8
7
6
5
4
0
1
2
3
4
Time

Simplest Analysis



- Forecast:
  Assume $P(X_{t+1}) \sim N(\mu_f, p_f)$

- Observation error:
  Assume $P(Y_{t+1} \mid X_{t+1}) \sim N(X_{t+1}, r)$

  - Likelihood = Data model

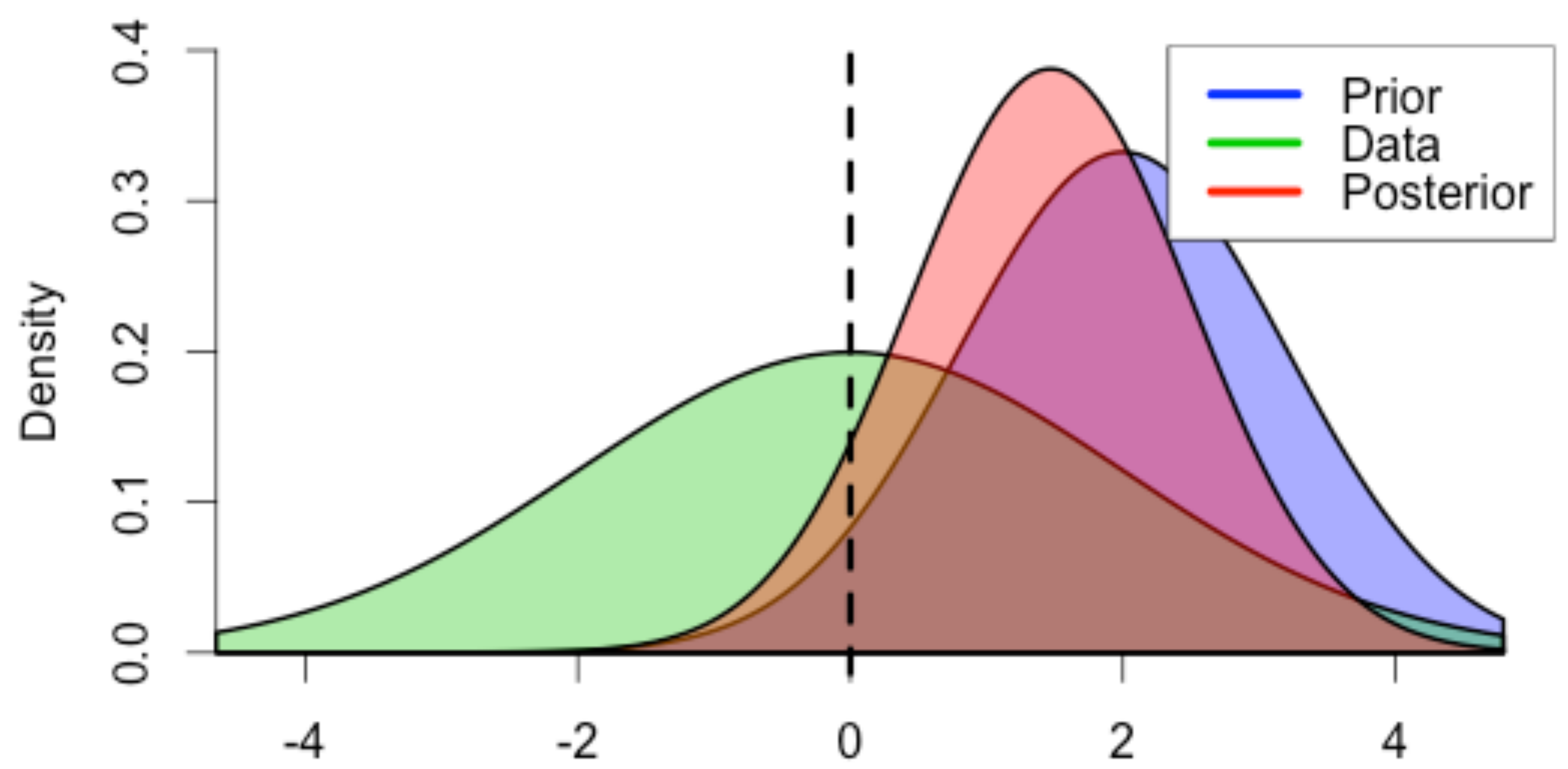
- Assume Y, μ_f, p_f and r are known

- $P(X_{t+1} \mid Y_{t+1}) \sim N(\mu_a, p_a)$

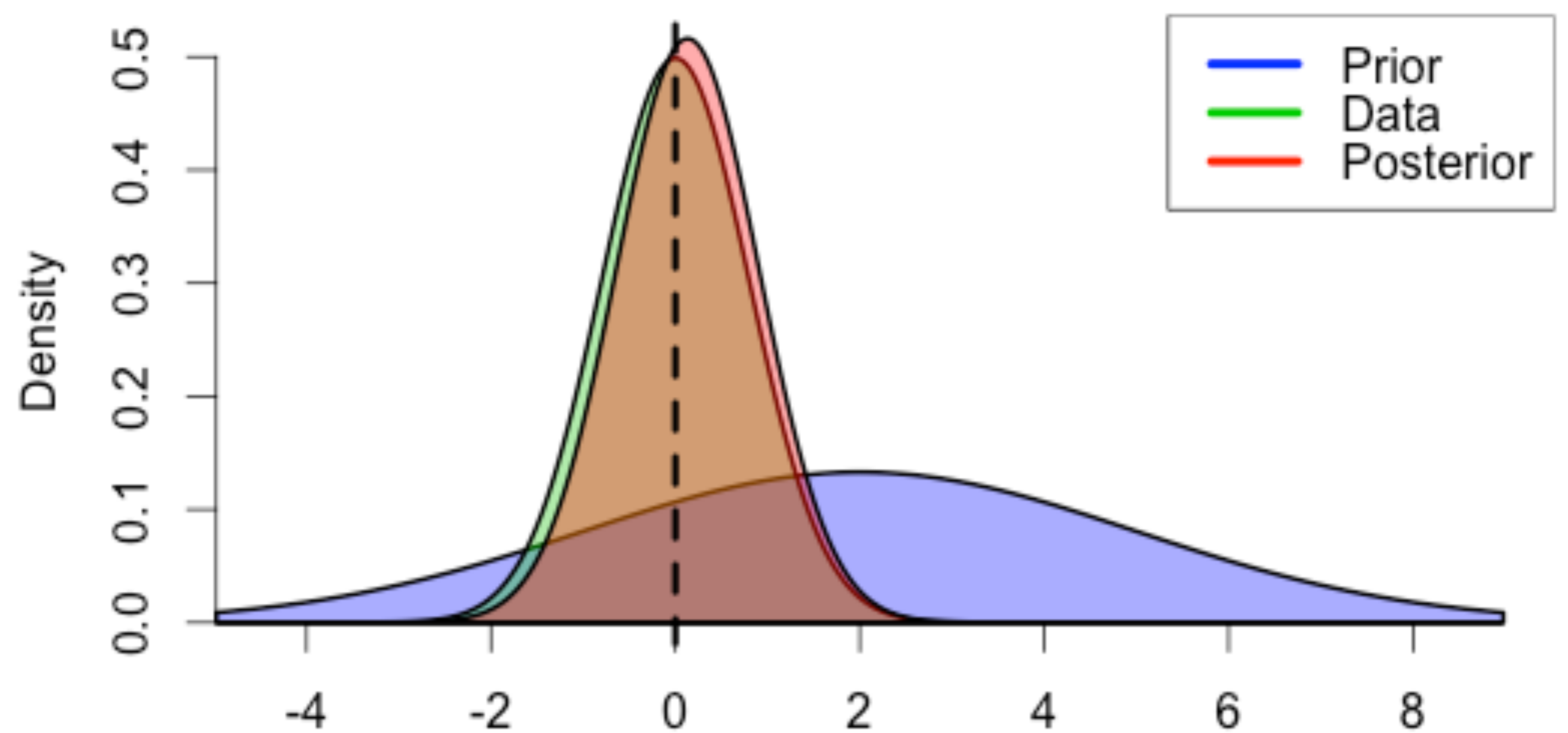
$$\rho = 1/r \qquad \phi = 1/p_f$$

$$X \mid Y \sim N\left(\frac{\rho}{n\rho + \phi} n\bar{Y} + \frac{\phi}{n\rho + \phi}\mu_f, n\rho + \phi\right)$$

Less Precise Data



Less Precise Model



Precision controls influence

Simplest Forecast

- Process Model
  $X_{t+1} = mX_t + \varepsilon_t$
- Process error
  $\varepsilon_t \sim N(0,q)$
- Assume m and q are known
- State uncertainty (IC)
  $P(X_t \mid Y_t) \sim N(\mu_a, p_a)$
- What is $P(X_{t+1})$?

- $E[X_{t+1}] = E[mX_t + \varepsilon_t] = m\mu_a$
- $Var[X_{t+1}] = Var[mX_t + \varepsilon_t]$
  - $m^2Var[X_t] + Var[\varepsilon_t]$ - $2Cov[mX_t,\varepsilon_t]$
  - $\approx m^2Var[X_t] + Var[\varepsilon_t]$
  - $m^2p_a + q$
- $P(X_{t+1}) \sim N(m\mu_a, m^2p_a + q)$

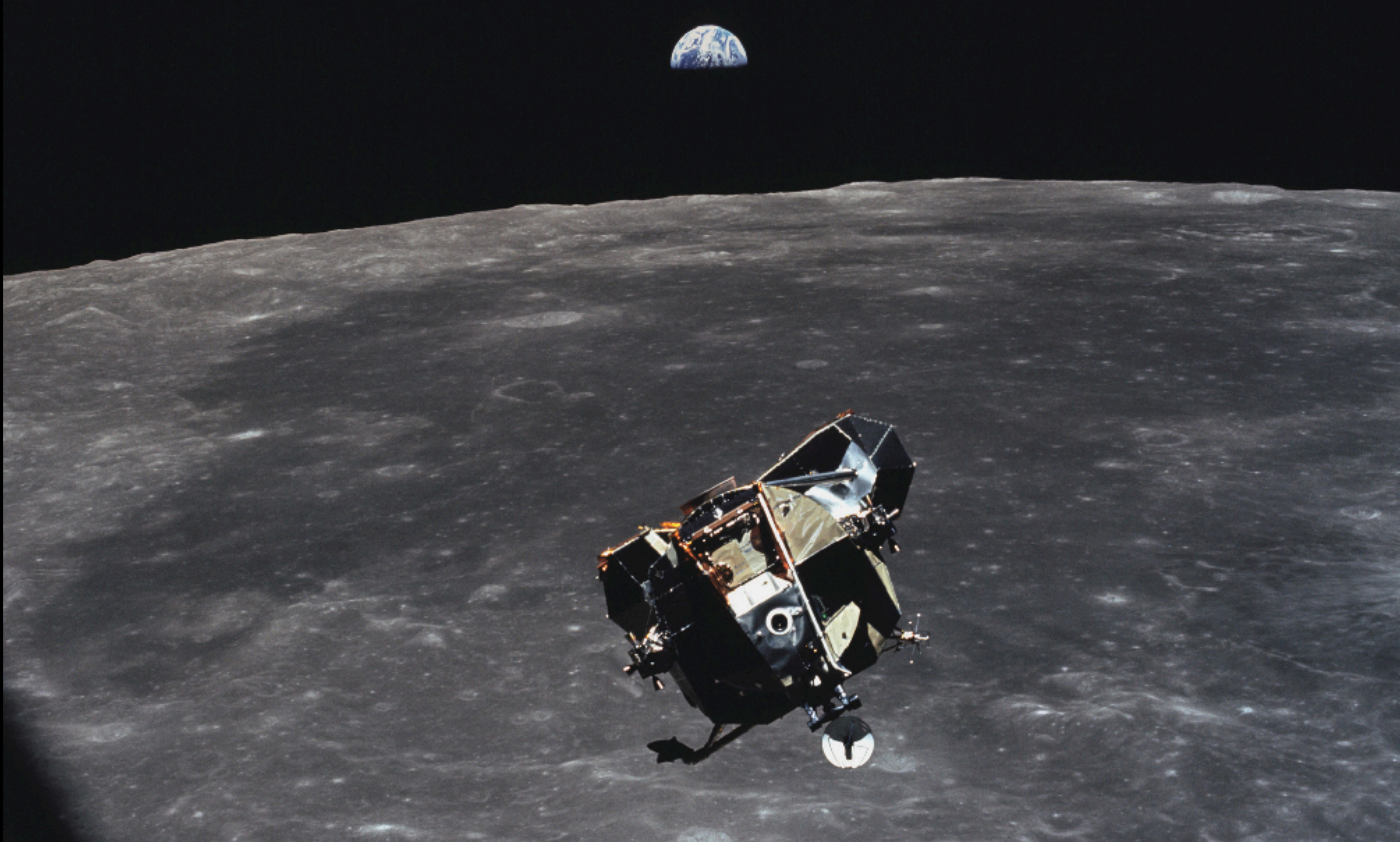
Forecast Cycle

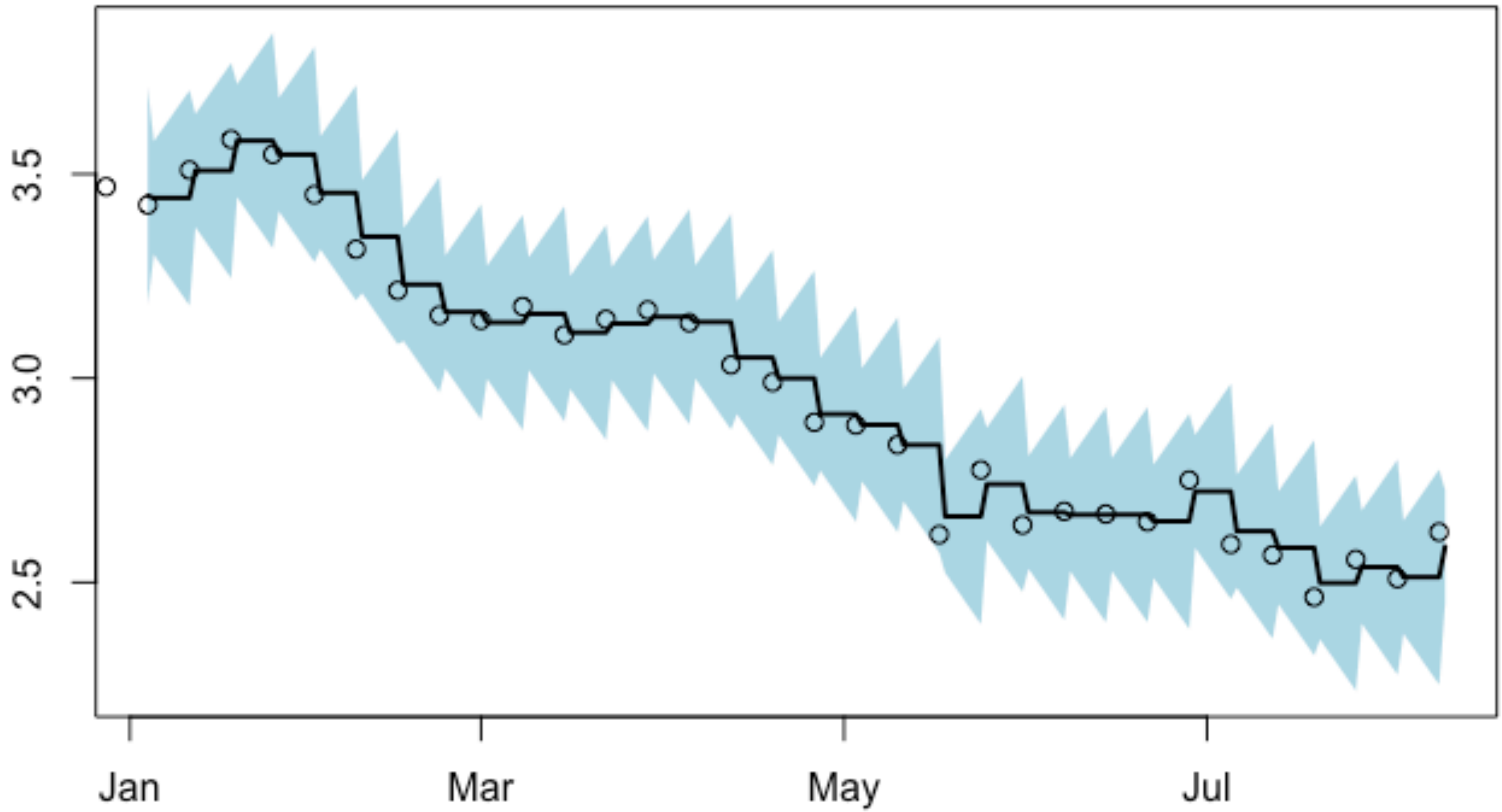
- Forecast Step:
  $P(X_{t+1}) \sim N(\mu_f = m\mu_a,$
  $p_f = m^2 p_a + q)$
- Analysis Step
  $P(X_{t+1} \mid Y_{t+1}) \sim N(\mu_a, p_a)$
  - $1/p_a = n/p_f + 1/r$
  - $\mu_a = (\mu_f/p_f + nY/r) \cdot p_a$
- Has an analytical solution!
- Kalman Filter

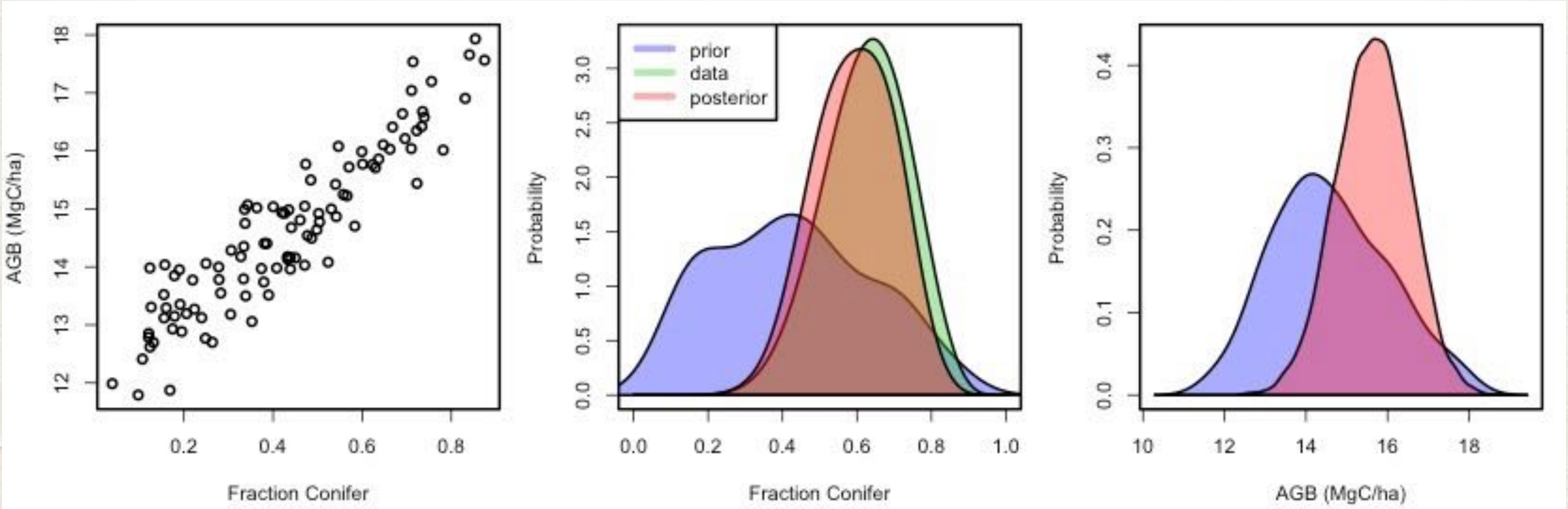

Rudolf Kalman

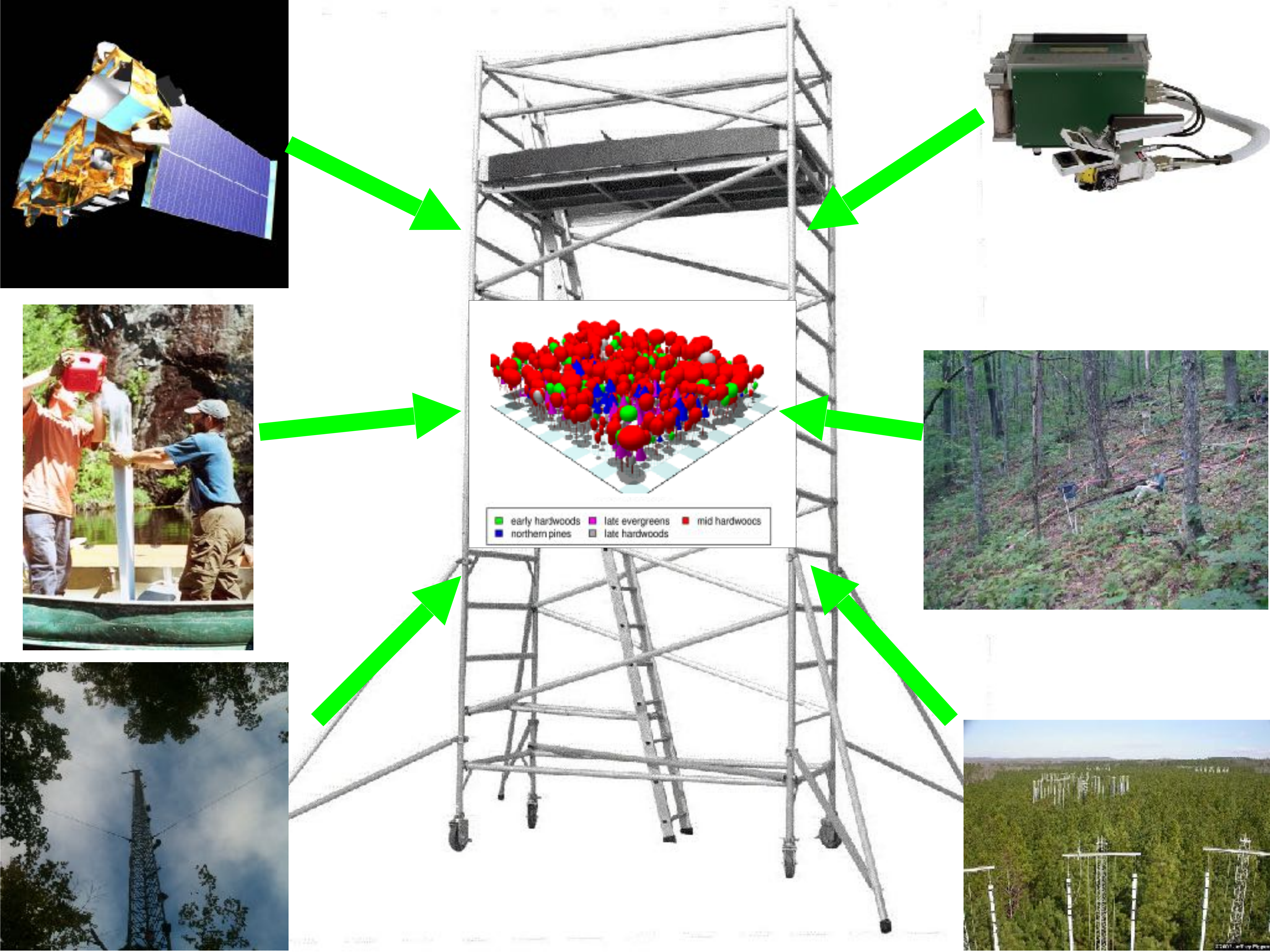
“Data assimilation isn’t rocket science, but you can use it for that.”

– DAVE MOORE




3.5
3.0
2.5
Jan
Mar
May
Jul


AGB (MgC/ha)
Fraction Conifer
Probability
prior
data
posterior
Fraction Conifer
Probability
AGB (MgC/ha)


early hardwoods
late evergreens
mid hardwoods
northern pines
late hardwoods

Generalized to Multivariate

- (n x 1) vector of state means, μ_a or μ_f
- (n x n) state error covariance matrix, P_a or P_f (was p_a, p_f)
- (p x 1) vector of observations, Y
- (p x p) observation error covariance matrix, R (was r)
- (p x n) observation matrix, H
- (n x n) linear process model, M (was m)
- (n x n) process error covariance matrix, Q (was q)

$$X_a|Y \sim N(Y|HX_a, R)\,N(X_a|\mu_f, P_f)$$

$$X_a|Y \sim N(Y|HX_a, R)\,N(X_a|\mu_f, P_f)$$

- Solves to be

$$X_a|Y \sim N\left(\left(H^T R^{-1} H + P_f^{-1}\right)^{-1}\left(H^T R^{-1} Y + P_f^{-1}\mu_f\right),\ \left(H^T R^{-1} H + P_f^{-1}\right)^{-1}\right)$$

$$P_a^{-1} = H^T R^{-1} H + P_f^{-1}$$

$$X_a | Y \sim N(Y | HX_a, R)\, N(X_a | \mu_f, P_f)$$

- Solves to be

$$X_a | Y \sim N\left(\left(H^T R^{-1} H + P_f^{-1}\right)^{-1} \left(H^T R^{-1} Y + P_f^{-1} \mu_f\right), \left(H^T R^{-1} H + P_f^{-1}\right)^{-1} \right)$$

- Mean and variance simplify to

$$E[X_a | Y] = \mu_a = \mu_f + K(Y - H\mu_f)$$

$$Var[X_a | Y] = P_a = (I - KH) P_f$$

$$K = P_{fH}^T \left(R + H P_f H^T\right)^{-1}$$ **Kalman Gain**

Example

- Assume $\mu_f = \{\mu_1,\mu_2,\mu_3\}$, $Y = \{y_2,y_3\}$, and observation error is $R = \sigma^2 I$

$$\begin{array}{cc} & \begin{array}{ccc} X_1 & X_2 & X_3 \end{array} \\ H = & \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{array} \begin{array}{l} Y_2 \\ Y_3 \end{array}$$

- The posterior mean for the unobserved X1 is

$$E[x_1] = \mu_1 + w_{12}(y_2 - \mu_2) + w_{13}(y_3 - \mu_3)$$

$$w_1^T = \underbrace{(p_{12} \quad p_{13})}_{\text{covariance between knowns and unknown}} \underbrace{\begin{pmatrix} p_{22}+\sigma^2 & p_{23} \\ p_{32} & p_{33}+\sigma^2 \end{pmatrix}^{-1}}_{\text{covariances among things we know}}$$

If X's are locations and P is a spatial covariance matrix, model is equivalent to Kriging

Forecast Step $X_{t+1} = MX_t + \epsilon$

The posterior distribution of X_{t+1} given X_t is multivariate normal with

$$\mu_{f,t+1} = E[X_{f,t+1} | X_{a,t}] = M_t \mu_{a,t}$$

$$P_{f,t+1} = Var[X_{f,t+1} | X_{a,t}] = Q_t + M_t P_{a,t-1} M_t^T$$

Pro/Con of Kalman Filter (KF)

- Analytically tractable
- Depends only upon the PREVIOUS state, the current Forecast, and the current Data

- Linear
- Normal
- Matrix inversion
- Assumes all parameters (H, R, M, Q) are **known**
- Forward only

UNCERTAINTY PROPAGATION
APPLIED IN THE FORECAST STEP

		Output	
		Distribution	Moments
Approach	Analytic	Variable Transform	Analytical Moments KF Taylor Series EKF
	Numeric	Monte Carlo PF	Ensemble EnKF

Extended Kalman Filter (EKF)

- Addresses **linear** assumption of the Forecast
  - $\mu_f = f(\mu_a)$
- Update variance using a Taylor Series expansion
  - F = Jacobian (df_i / dx_j)
  - $P_f \approx Q + F P_a F^T$ (was $Q + M P_a M^T$)
  - Can be extended to higher orders
- Jensen's Inequality: Biased, Normality assumption FALSE

$$N_{t+1} = N_t + rN_t\left(1 + \frac{N}{K}\right) \quad \longrightarrow \quad \frac{\partial N_{t+1}}{\partial N_t} = 1 + r - \frac{2r}{K}N_t$$