Astrostatistics: 06 Mar 2019

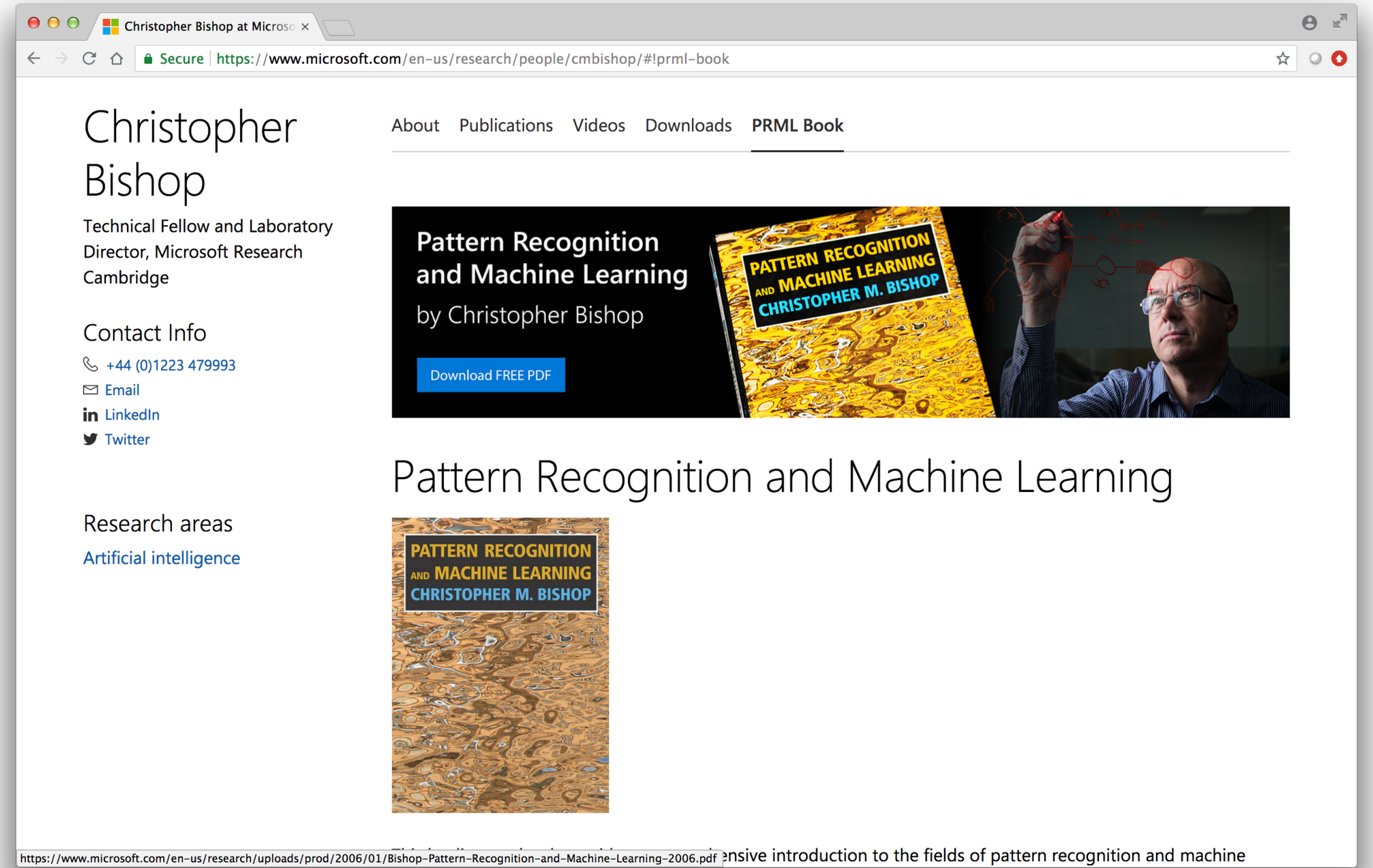
https://github.com/CambridgeAstroStat/PartIII-Astrostatistics-2019

- Example Class 3: Thu, 14 March, 12-2pm MR9
- Example Class 4: Probably ~ April 29-30
- Make-up Lecture:
  - Tue 12 Mar, 2pm MR 14, or Fri 15 Mar 12pm MR13 ?
- Finish Gaussian Processes in Astrophysics
- Case Study: application to gravitationally lensed supernova
- Hierarchical Bayes

Human Learning of Gaussian Processes

- Classic Text: Rasmussen & Williams (2006)
  - “Gaussian Processes for Machine Learning”,Ch 1-2,4-5
  - Free Online: http://www.gaussianprocess.org/gpml/
- Ivezic, Sec 8.10 GP Regression, (Ch 8 is Regression)
- **Bishop: Pattern Recognition & Machine Learning, Ch 6**
  - **Also free online: https://www.microsoft.com/en-us/research/people/cmbishop/#!prml-book**
- Gelman, Bayesian Data Analysis 3rd Ed., Chapter 21
- “Practical Introduction to GPs for Astronomy” - D. Foreman-Mackey
  - http://hea-www.harvard.edu/AstroStat/aas231_2018/DForeman-Mackey_20180110_aas231.pdf

Human Learning of Gaussian Processes



Review: Fitting a GP to data

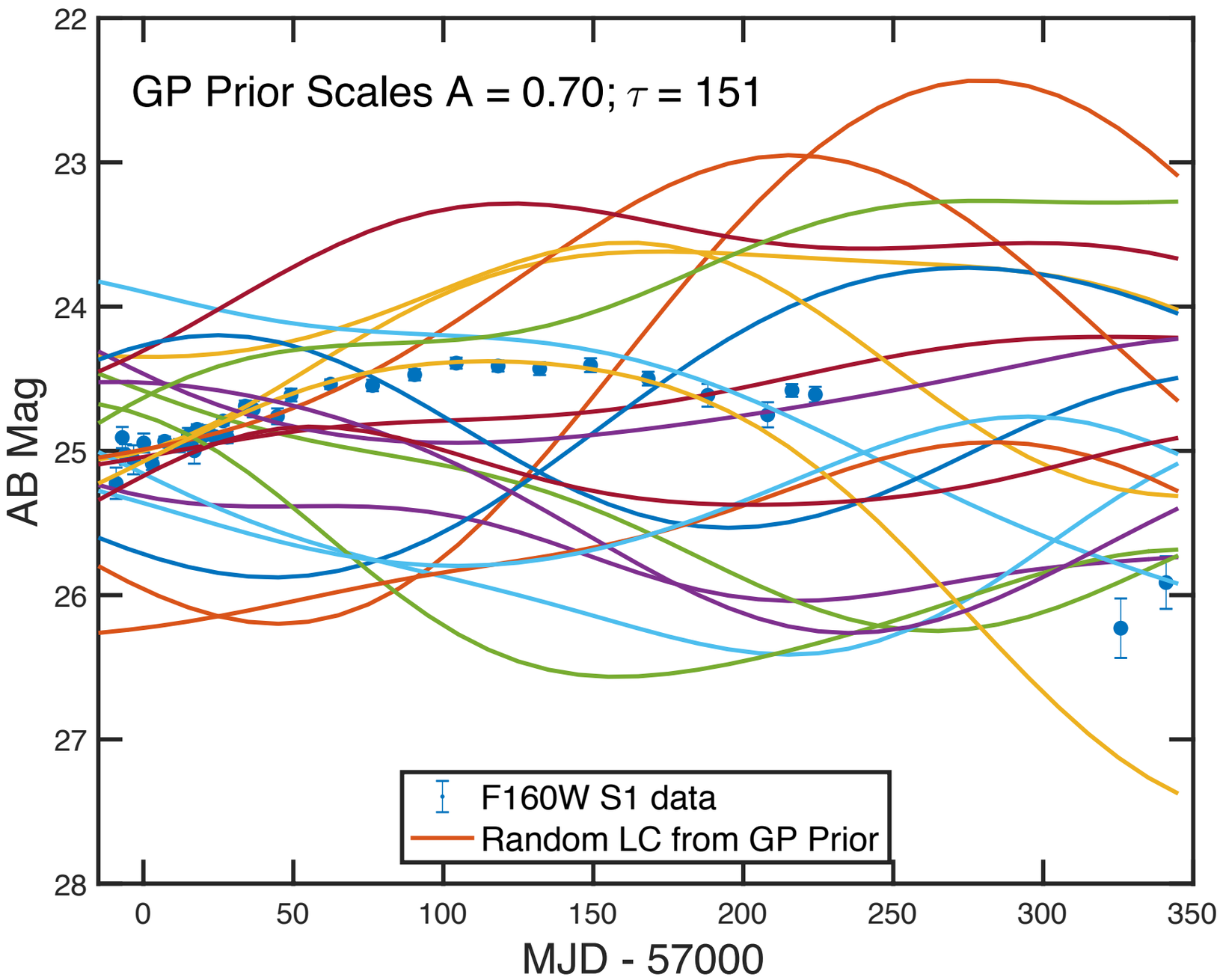
1. If we knew the characteristic scales of the kernel (A, $\boldsymbol{\tau}^2$), then how do we fit the data at observed times to find the curve for unobserved times? (computing the posterior)

2. How do we fit for the characteristic scales of the kernel (hyperparameters)? (model selection)

1. Which curve from the prior is the best description of the data?

$$\mathrm{Cov}[f(t), f(t')] = k(t, t') = A \exp(-|t - t'|^2/\tau^2)$$

$$\boldsymbol{f} | \, A, \tau \sim N(\mathbf{1}c, \boldsymbol{K})$$



Posterior Inference with GPs
Estimating the underlying curve:

$\boldsymbol{f}_o = \text{observed points at times } \boldsymbol{t}_o$ (training set)

$\boldsymbol{f}_* = \text{function at unobserved times } \boldsymbol{t}_*$ (prediction or test set)

Joint:
$$\begin{pmatrix} \boldsymbol{f}_o \\ \boldsymbol{f}_* \end{pmatrix} \sim N\left(\begin{bmatrix} \mathbf{1}c \\ \mathbf{1}c \end{bmatrix}, \begin{bmatrix} \boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_o) & \boldsymbol{K}(\boldsymbol{t}_*, \boldsymbol{t}_o) \\ \boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_*) & \boldsymbol{K}(\boldsymbol{t}_*, \boldsymbol{t}_*) \end{bmatrix} \right)$$

Populating the Covariance Matrix

$$\boldsymbol{K}(\boldsymbol{t}, \boldsymbol{t}') \text{ has i,j-th entry} = k(t_i, t'_j)$$

Using the assumed kernel function

$$\text{Cov}[f(t), f(t')] = k(t, t') = A \exp(-|t - t'|^2/\tau^2)$$

Review: Posterior Inference with GPs

Estimating the underlying curve:

$\boldsymbol{f}_o = \text{observed points at times } \boldsymbol{t}_o$

$\boldsymbol{f}_* = \text{function at unobserved times } \boldsymbol{t}_*$

Jointly Gaussian: $\begin{pmatrix} \boldsymbol{f}_o \\ \boldsymbol{f}_* \end{pmatrix} \sim N\left(\begin{bmatrix} \mathbf{1}c \\ \mathbf{1}c \end{bmatrix}, \begin{bmatrix} \boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_o) & \boldsymbol{K}(\boldsymbol{t}_*, \boldsymbol{t}_o) \\ \boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_*) & \boldsymbol{K}(\boldsymbol{t}_*, \boldsymbol{t}_*) \end{bmatrix} \right)$

Posterior is also Gaussian $\boldsymbol{f}_* | \boldsymbol{f}_o \sim N\left(\mathbb{E}[\boldsymbol{f}_* | \boldsymbol{f}_o], \text{Var}[\boldsymbol{f}_* | \boldsymbol{f}_o]\right)$

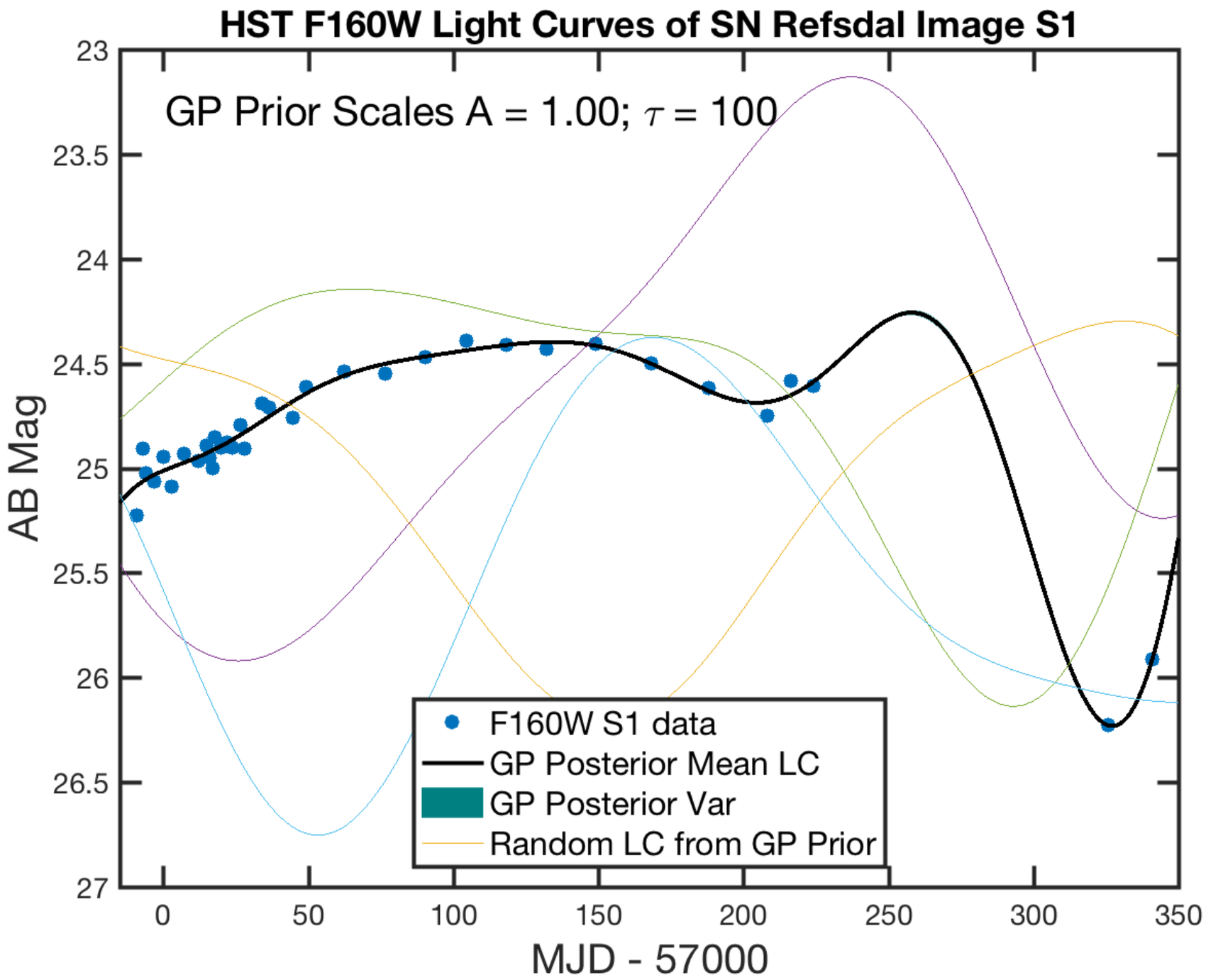
Posterior Mean:

$$\mathbb{E}[\boldsymbol{f}_* | \boldsymbol{f}_o] = \mathbf{1}c + \boldsymbol{K}(\boldsymbol{t}_*, \boldsymbol{t}_o)\boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_o)^{-1}(\boldsymbol{f}_o - \mathbf{1}c)$$

Posterior Co(variance):

$$\text{Var}[\boldsymbol{f}_* | \, \boldsymbol{f}_o] = \boldsymbol{K}(\boldsymbol{t}_*, \boldsymbol{t}_*) - \boldsymbol{K}(\boldsymbol{t}_*, \boldsymbol{t}_o)\boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_o)^{-1}\boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_*)$$

Posterior Inference with GPs
Estimating the underlying curve:



Accounting for Measurement Error

$$\boldsymbol{y}_o | \boldsymbol{f}_o \sim N(\boldsymbol{f}_o, \boldsymbol{W})$$

$\boldsymbol{y}_o$ are measured values of $\boldsymbol{f}_o$ at times $\boldsymbol{t}_o$

W is measurement covariance matrix
(often diagonal for independent noise)

$$W_{ij} = \delta_{ij}\sigma_i^2$$

$$\begin{pmatrix} \boldsymbol{y}_o \\ \boldsymbol{f}_* \end{pmatrix} \sim N\left(\begin{bmatrix} \mathbf{1}c \\ \mathbf{1}c \end{bmatrix}, \begin{bmatrix} \boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_o) + \boldsymbol{W} & \boldsymbol{K}(\boldsymbol{t}_*, \boldsymbol{t}_o) \\ \boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_*) & \boldsymbol{K}(\boldsymbol{t}_*, \boldsymbol{t}_*) \end{bmatrix} \right)$$

Now can calculate function prediction at unobserved points

$$\boldsymbol{f}_* | \boldsymbol{y}_o \sim N(\mathbb{E}[\boldsymbol{f}_* | \boldsymbol{y}_o], \mathrm{Var}[\boldsymbol{f}_* | \boldsymbol{y}_o])$$

Using conditional properties of Gaussian as before

Accounting for Measurement Error: Derivation as the sum of two GPs at the observed times

GP of Intrinsic Curve

$$f(t) \sim \mathcal{GP}(m(t) = c, k(t, t'))$$

$$\boldsymbol{f}_o = \text{ function at observed times } \boldsymbol{t}_o$$

$$\boldsymbol{f}_o \sim N[\mathbf{1}_c, \boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_o)]$$

GP of Measurement Error

$$\boldsymbol{y}_o | \boldsymbol{f}_o \sim N(\boldsymbol{f}_o, \boldsymbol{W})$$

Same as: (mean-zero noise)

$$\boldsymbol{y}_o = \boldsymbol{f}_o + \boldsymbol{\epsilon} \qquad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{W})$$

Most common case:
heteroskedastic uncorrelated measurement error:

$$\text{Cov}(\epsilon_i, \epsilon_j) \equiv W_{ij} = \delta_{ij}\, \sigma_i^2$$

Accounting for Measurement Error:
Derivation as the sum of two GPs at the observed times

Intrinsic/Latent Process: $\boldsymbol{f}_o \sim N[\mathbf{1}_c, \boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_o)]$

Measurement Process: $\boldsymbol{y}_o|\boldsymbol{f}_o \sim N(\boldsymbol{f}_o, \boldsymbol{W})$

$$\boldsymbol{y}_o = \boldsymbol{f}_o + \boldsymbol{\epsilon} \qquad\qquad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{W})$$

$$\mathrm{Cov}(\boldsymbol{y}_o, \boldsymbol{y}_o) = \mathrm{Cov}(\boldsymbol{f}_o, \boldsymbol{f}_o) + \mathrm{Cov}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}) + 2\,\mathrm{Cov}(\boldsymbol{f_o}, \boldsymbol{\epsilon})$$

$\mathrm{Cov}(\boldsymbol{f}_o, \boldsymbol{f}_o) = \boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_o)$ (GP of intrinsic curve)

$\mathrm{Cov}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}) = \boldsymbol{W}$ (measurement noise)

(the two processes are uncorrelated)

$$2\mathrm{Cov}[\boldsymbol{f}_o, \boldsymbol{\epsilon}] = 0$$

Therefore: $\mathrm{Cov}[\boldsymbol{y}_o, \boldsymbol{y}_o] = \boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_o) + \boldsymbol{W}$

Accounting for Measurement Error: Derivation as the sum of two GPs at the observed times

$$\boldsymbol{y}_o|\boldsymbol{f}_o \sim N(\boldsymbol{f}_o, \boldsymbol{W})$$

$$\boldsymbol{y}_o = \boldsymbol{f}_o + \boldsymbol{\epsilon} \qquad\qquad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{W})$$

$$\begin{pmatrix} \boldsymbol{f}_o \\ \boldsymbol{f}_* \end{pmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{1}c \\ \boldsymbol{1}c \end{bmatrix}, \begin{bmatrix} \boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_o) & \boldsymbol{K}(\boldsymbol{t}_*, \boldsymbol{t}_o) \\ \boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_*) & \boldsymbol{K}(\boldsymbol{t}_*, \boldsymbol{t}_*) \end{bmatrix} \right)$$

Similar arguments for:

$$\mathrm{Cov}[\boldsymbol{y}_o, \boldsymbol{f}_*] = \mathrm{Cov}[\boldsymbol{f}_o, \boldsymbol{f}_*] + \mathrm{Cov}[\boldsymbol{\epsilon}, \boldsymbol{f}_*]$$

$$\mathrm{Cov}[\boldsymbol{y}_o, \boldsymbol{f}_*] = \boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_*) + \boldsymbol{0} = \boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_*)$$

$$\begin{pmatrix} \boldsymbol{y}_o \\ \boldsymbol{f}_* \end{pmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{1}c \\ \boldsymbol{1}_c \end{bmatrix}, \begin{bmatrix} \boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_o) + \boldsymbol{W} & \boldsymbol{K}(\boldsymbol{t}_*, \boldsymbol{t}_o) \\ \boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_*) & \boldsymbol{K}(\boldsymbol{t}_*, \boldsymbol{t}_*) \end{bmatrix} \right)$$

Accounting for Measurement Error:
Derivation using Conditional/Marginal properties of MV Gaussian

$$\begin{pmatrix} \boldsymbol{f}_o \\ \boldsymbol{f}_* \end{pmatrix} \sim N\left(\begin{bmatrix} \mathbf{1}c \\ \mathbf{1}c \end{bmatrix}, \begin{bmatrix} \boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_o) & \boldsymbol{K}(\boldsymbol{t}_*, \boldsymbol{t}_o) \\ \boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_*) & \boldsymbol{K}(\boldsymbol{t}_*, \boldsymbol{t}_*) \end{bmatrix} \right)$$

$$\begin{pmatrix} \boldsymbol{y}_o \\ \boldsymbol{f}_* \end{pmatrix} \Big| \begin{pmatrix} \boldsymbol{f}_o \\ \boldsymbol{f}_* \end{pmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{f}_o \\ \boldsymbol{f}_* \end{bmatrix}, \begin{bmatrix} \boldsymbol{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \right)$$

P(V) & P(U|V) —> P(U)

$$\begin{pmatrix} \boldsymbol{y}_o \\ \boldsymbol{f}_* \end{pmatrix} \sim N\left(\begin{bmatrix} \mathbf{1}c \\ \mathbf{1}_c \end{bmatrix}, \begin{bmatrix} \boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_o) + \boldsymbol{W} & \boldsymbol{K}(\boldsymbol{t}_*, \boldsymbol{t}_o) \\ \boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_*) & \boldsymbol{K}(\boldsymbol{t}_*, \boldsymbol{t}_*) \end{bmatrix} \right)$$

Now can calculate function prediction at unobserved points

$$\boldsymbol{f}_* | \boldsymbol{y}_o \sim N(\mathbb{E}[\boldsymbol{f}_* | \boldsymbol{y}_o], \mathrm{Var}[\boldsymbol{f}_* | \boldsymbol{y}_o])$$

Using conditional properties of Gaussian as before

Accounting for Measurement Error:

$$\begin{pmatrix} \boldsymbol{y}_o \\ \boldsymbol{f}_* \end{pmatrix} \sim N\left(\begin{bmatrix} \mathbf{1}c \\ \mathbf{1}c \end{bmatrix}, \begin{bmatrix} \boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_o) + \boldsymbol{W} & \boldsymbol{K}(\boldsymbol{t}_*, \boldsymbol{t}_o) \\ \boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_*) & \boldsymbol{K}(\boldsymbol{t}_*, \boldsymbol{t}_*) \end{bmatrix} \right)$$

Now can calculate function prediction at unobserved points

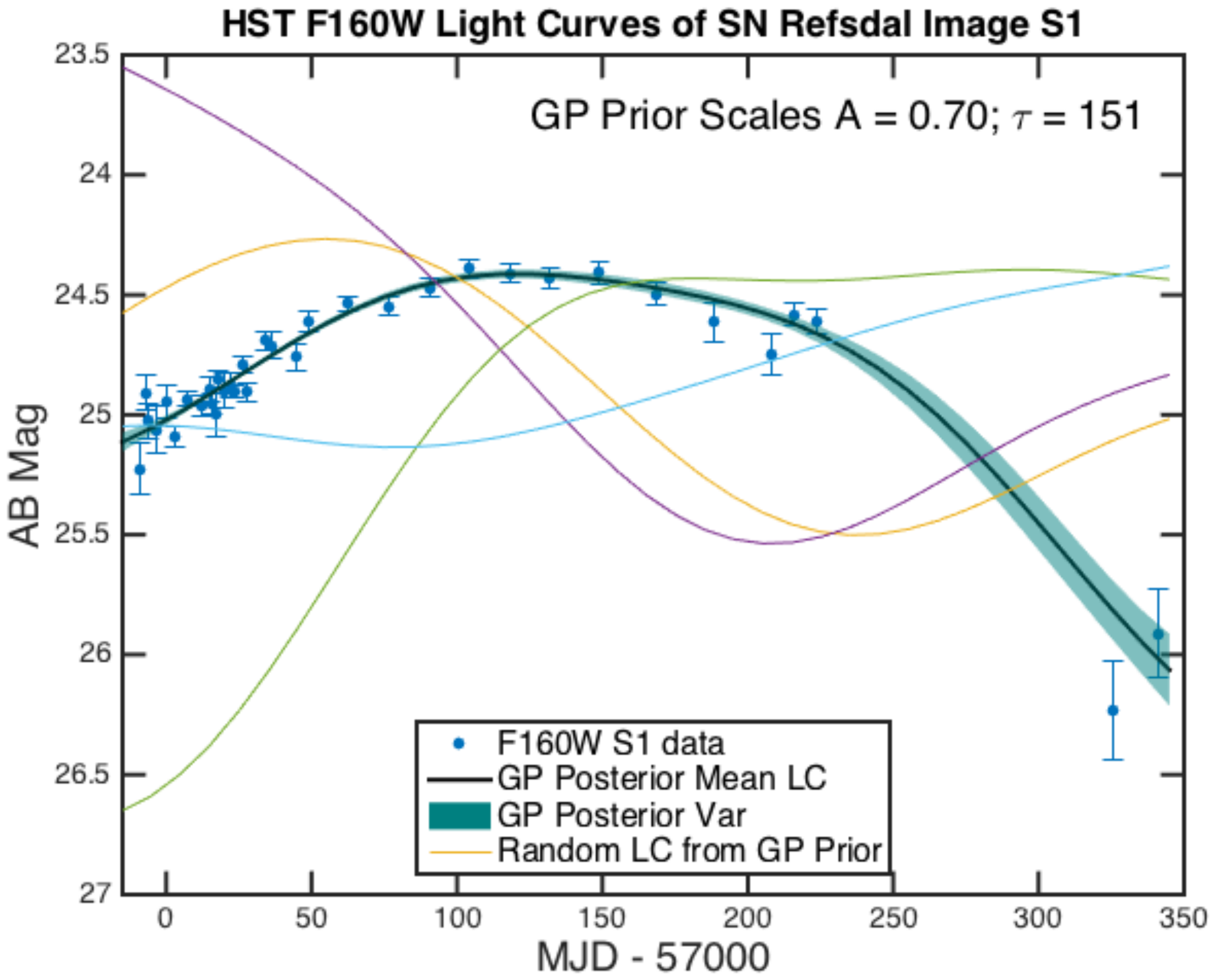
$$\boldsymbol{f}_* | \boldsymbol{y}_o \sim N(\mathbb{E}[\boldsymbol{f}_* | \boldsymbol{y}_o], \mathrm{Var}[\boldsymbol{f}_* | \boldsymbol{y}_o])$$

Using Gaussian Conditional Properties:

$$\mathbb{E}[\boldsymbol{f}_* | \boldsymbol{y}_o] = \mathbf{1}c + \boldsymbol{K}(\boldsymbol{t}_*, \boldsymbol{t}_o)[\boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_o) + \boldsymbol{W}]^{-1}(\boldsymbol{y}_o - \mathbf{1}c)$$

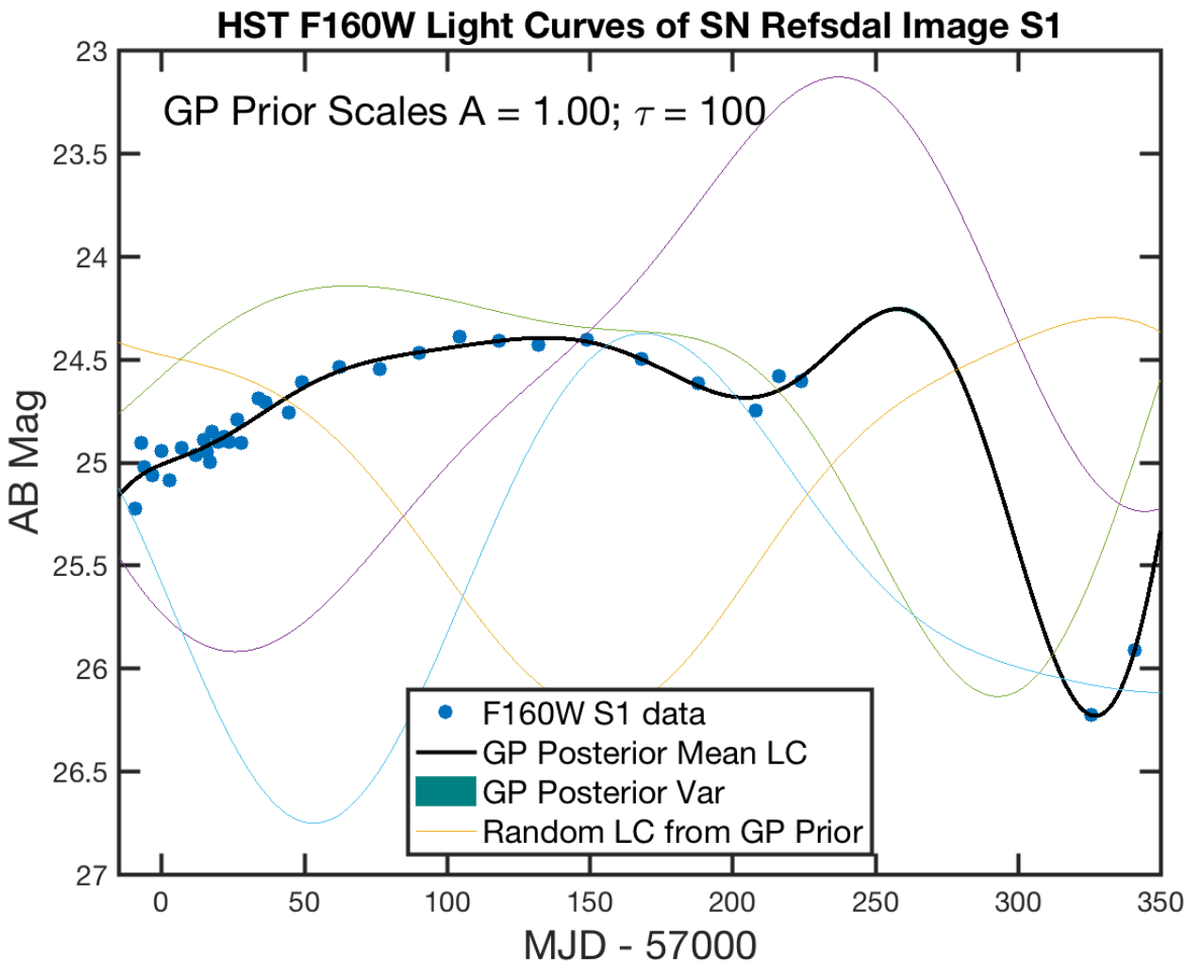
$$\mathrm{Var}[\boldsymbol{f}_* | \boldsymbol{y}_o] = \boldsymbol{K}(\boldsymbol{t}_*, \boldsymbol{t}_*) - \boldsymbol{K}(\boldsymbol{t}_*, \boldsymbol{t}_o)[\boldsymbol{K}(\boldsymbol{t_o}, \boldsymbol{t}_o) + \boldsymbol{W}]^{-1}\boldsymbol{K}(\boldsymbol{t}, \boldsymbol{t}_*)$$

Posterior Inference with GPs Accounting for Measurement Error



GP Fitting with measurement error

(over)fitting f_0

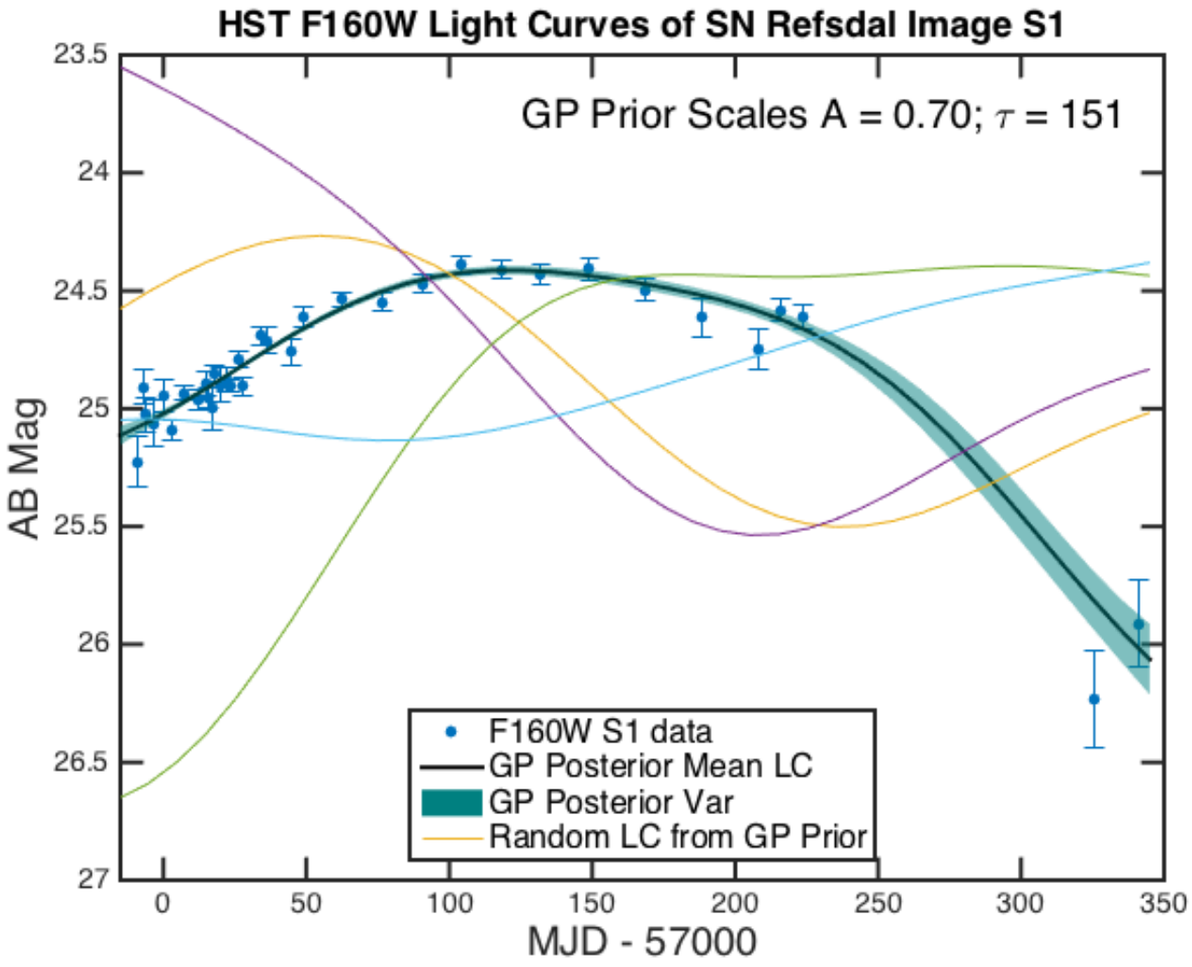


Actually a small "nugget term" was added to regularise matrix inversion - acts like meas. err.

$$\boldsymbol{f}_0 \sim N(\mathbf{1}c, \boldsymbol{K} + \sigma^2 \boldsymbol{I})$$

$$\sigma = 10^{-3}$$

Fitting y_0

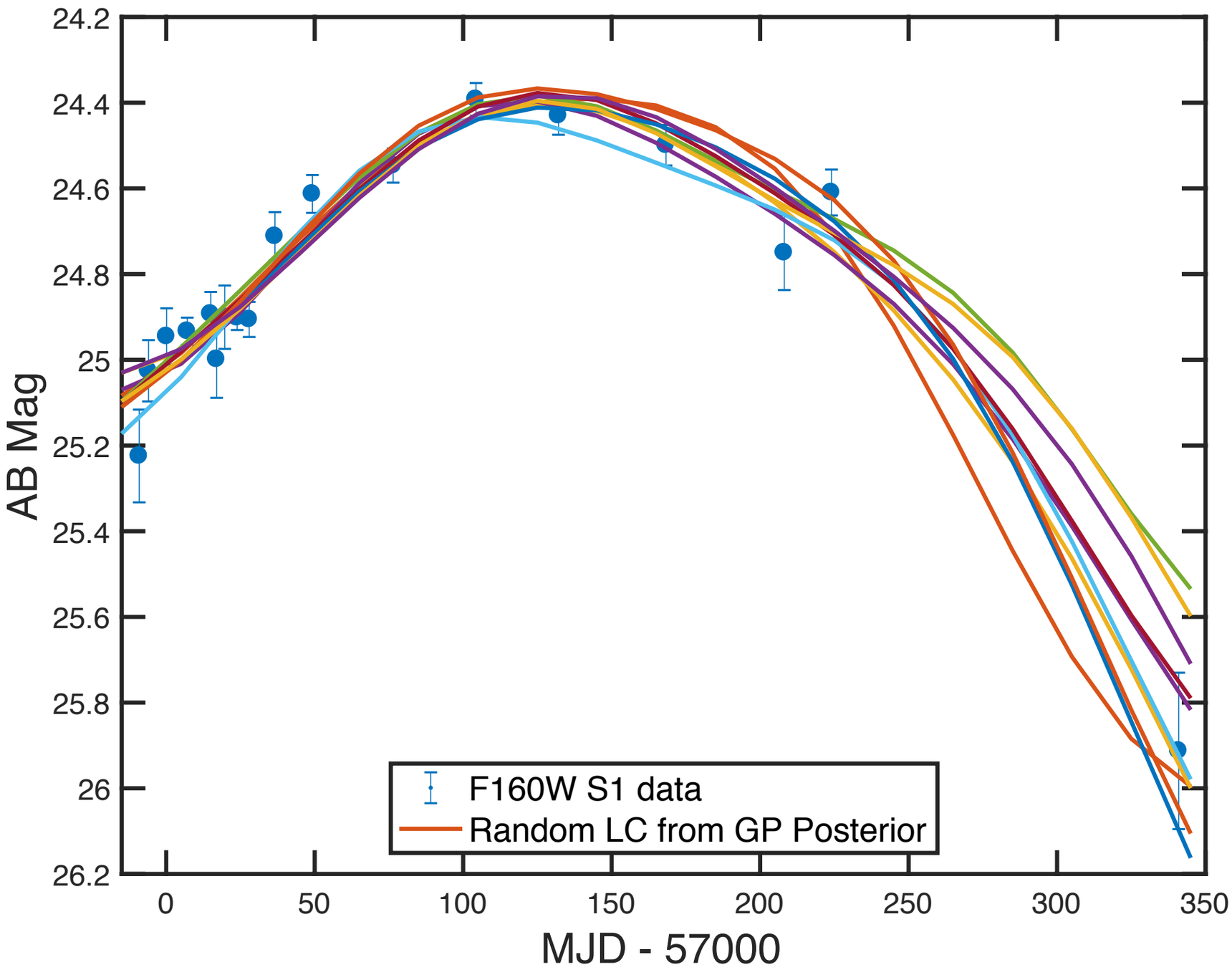


Using actual measurement uncertainties

$$\boldsymbol{y}_0 \sim N(\mathbf{1}c, \boldsymbol{K} + \boldsymbol{W})$$

$$W_{ij} = \sigma_i^2 \, \delta_{ij}$$

Posterior Inference with GPs
Accounting for Measurement Error:
Random draws from the posterior given noisy data



Fitting a GP to data

1. If we knew the characteristic scales of the kernel (A, $\boldsymbol{\tau}^2$), then how do we fit the data at observed times to find the curve for unobserved times? (computing the posterior)

2. How do we fit for the characteristic scales of the kernel (hyperparameters)? (model selection)

Bayesian Model Selection: tuning the hyperparameters (A, $\boldsymbol{\tau}$)

Integrating out the latent function f(t)
gives us the Marginal Likelihood:

$$P(\boldsymbol{y}_o|\, A, \tau^2) = \int P(\boldsymbol{y}_o|\, \boldsymbol{f}_o) \times P(\boldsymbol{f}_o|\, A, \tau^2)\, d\boldsymbol{f}_o$$

$$P(\boldsymbol{y}_o|\, A, \tau^2) = \int N[\boldsymbol{y}_o|\, \boldsymbol{f}_o, \boldsymbol{W}] \times N[\boldsymbol{f}_o|\, \mathbf{1}c, \boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_o)]\, d\boldsymbol{f}_o$$

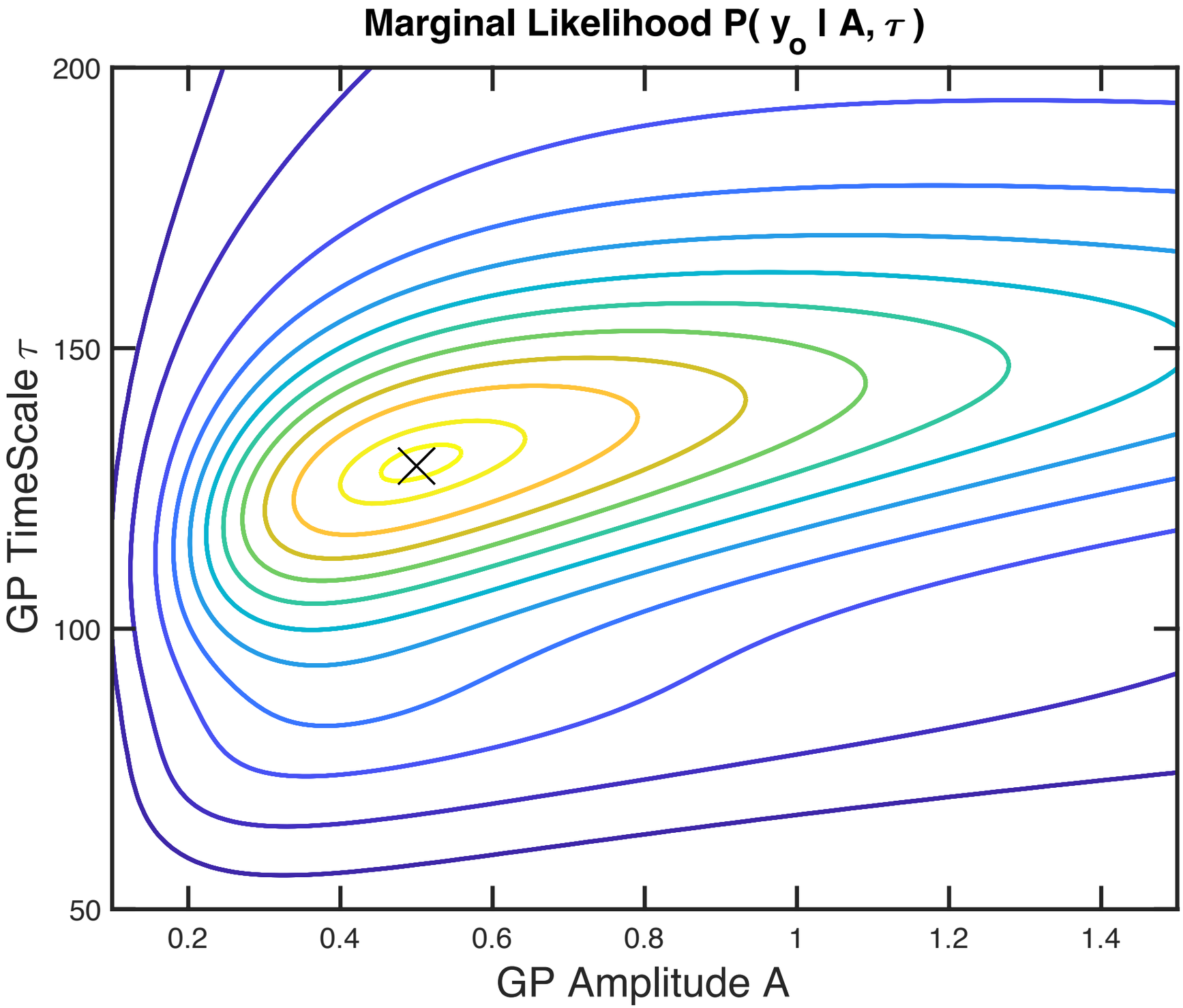
$$L(A, \tau^2) = P(\boldsymbol{y}_o|A, \tau^2) = N[\boldsymbol{y_o}|\, \mathbf{1}c, \boldsymbol{K}_{A,\tau^2}(\boldsymbol{t}_o, \boldsymbol{t}_o) + \boldsymbol{W}]$$

$$\boldsymbol{W} = \text{ measurement error covariance}$$

$$\boldsymbol{K}_{A,\tau^2} = \text{ GP covariance}$$

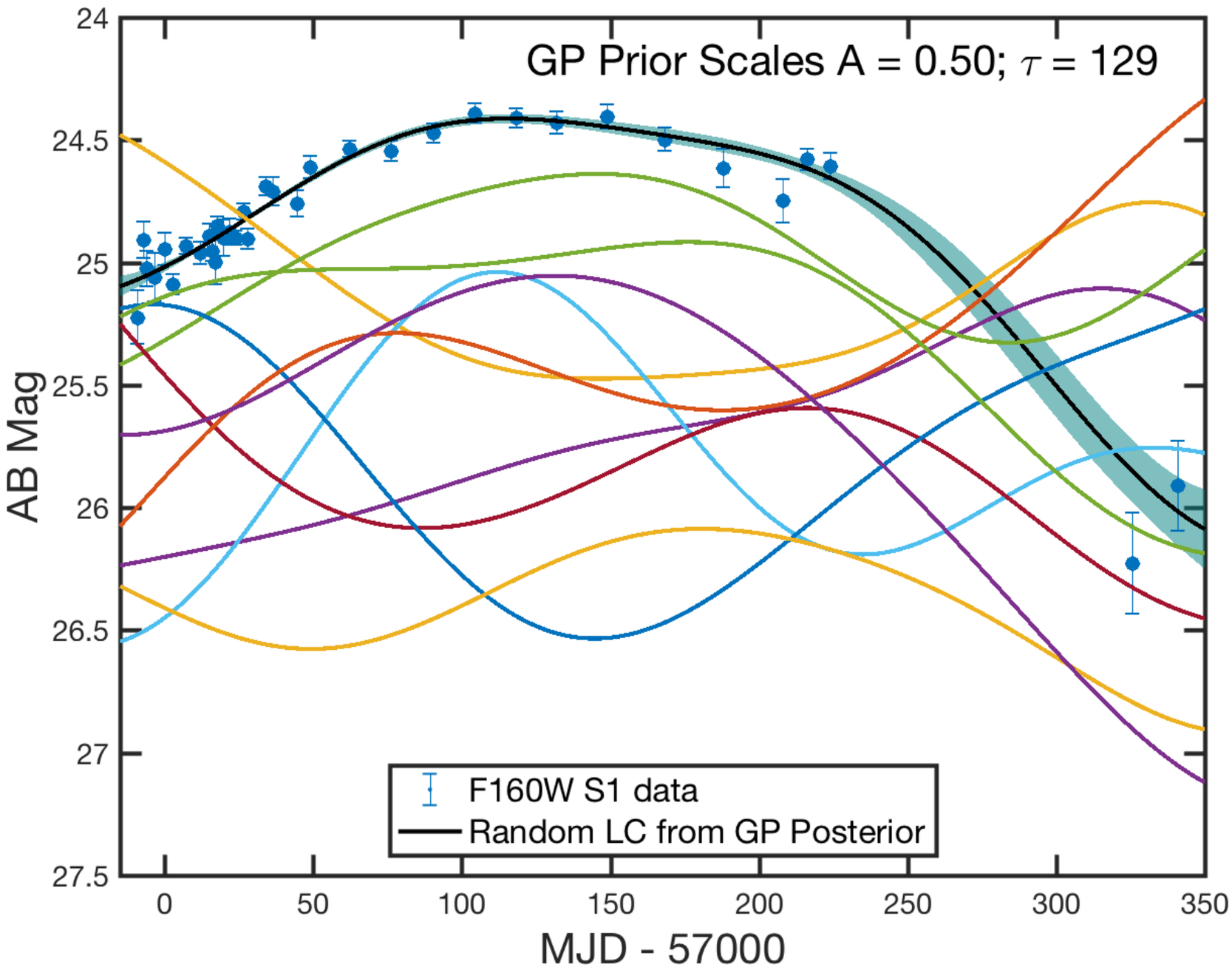
Which we can optimise (max likelihood)
or specify a prior on (A, $\boldsymbol{\tau}$) and sample from posterior

Demonstration

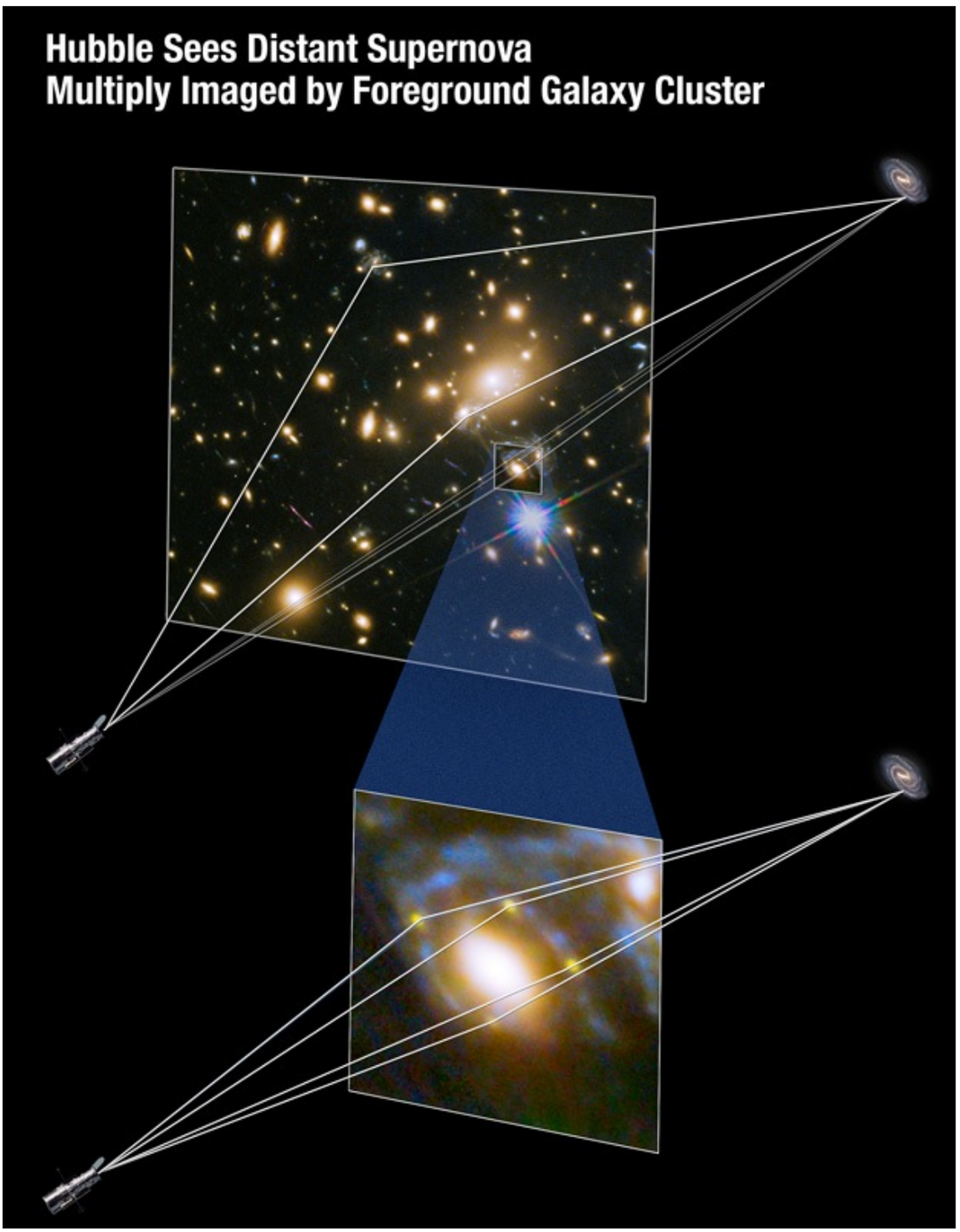


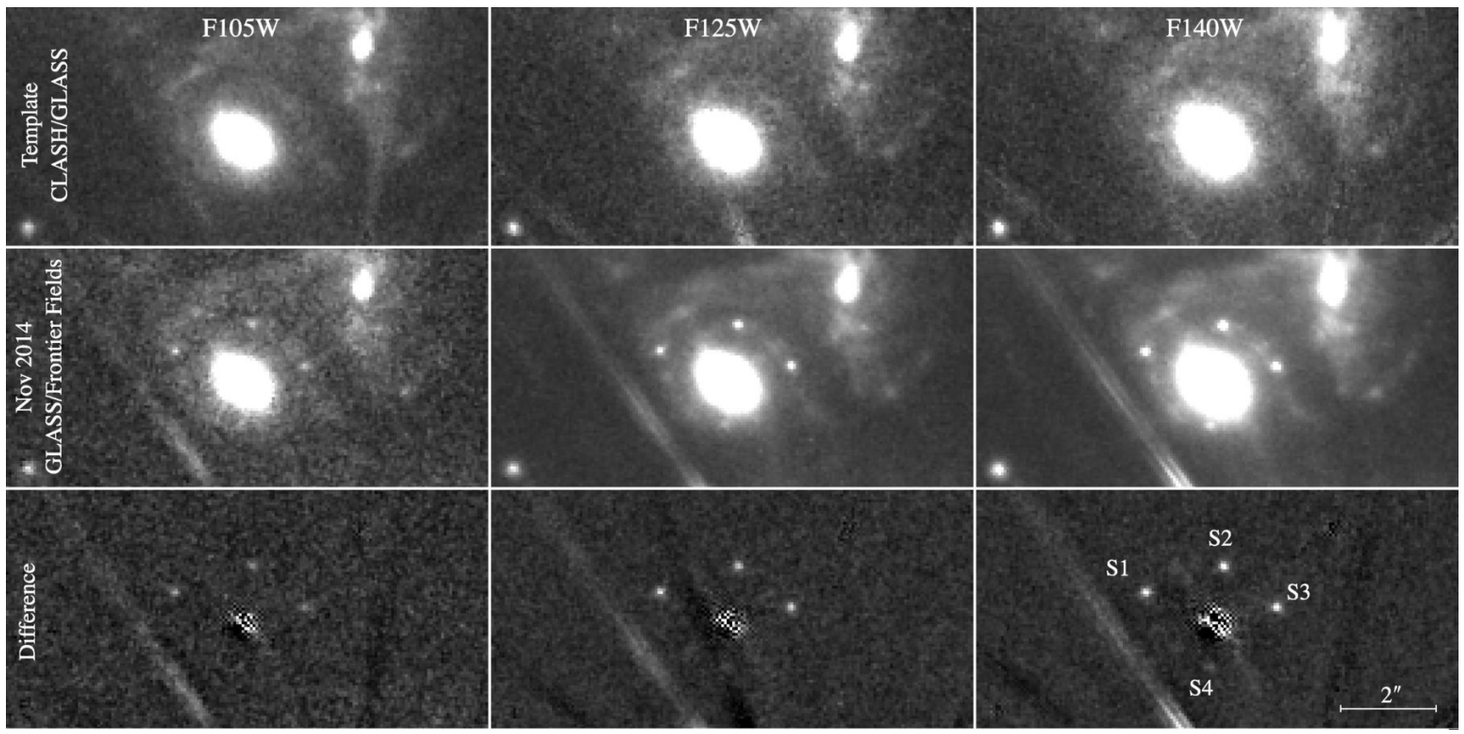
MLE: A = 0.50, $\boldsymbol{\tau}$ = 129

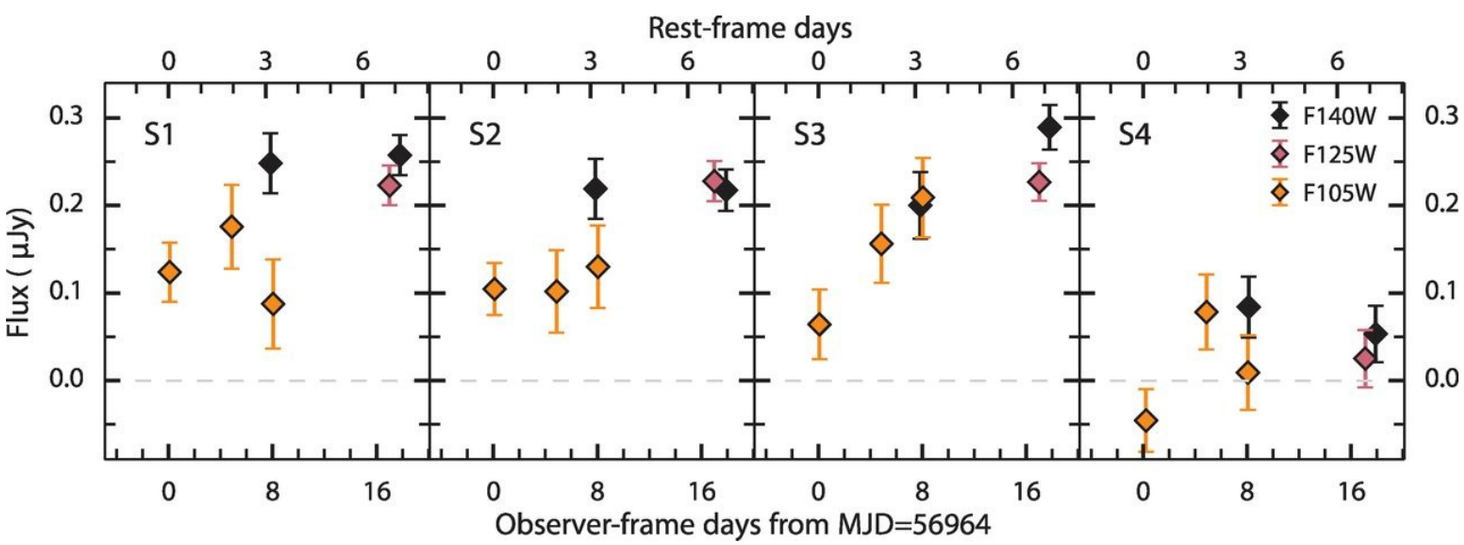
GP Fit with optimised hyperparameters



Case Study: SN Refsdal

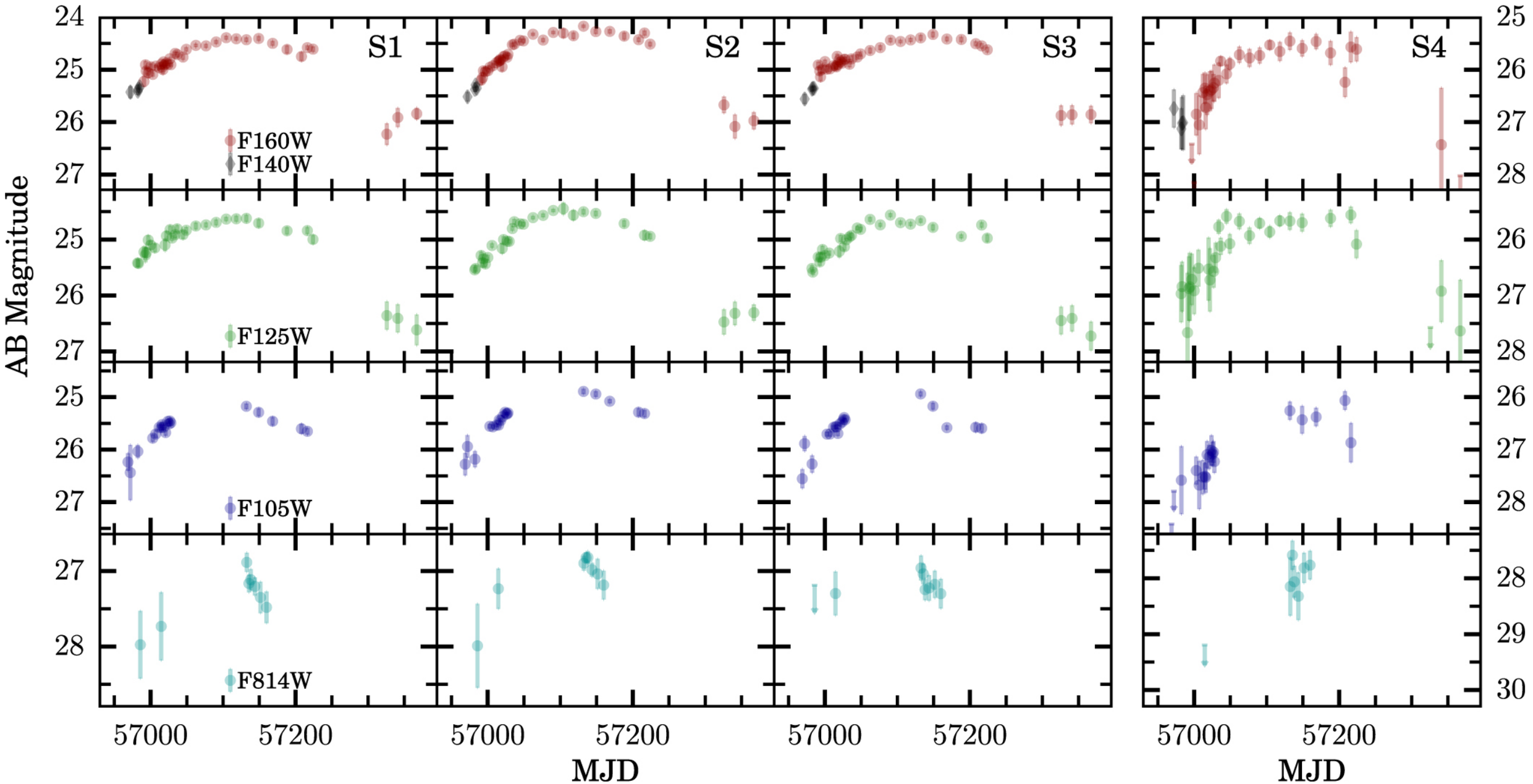






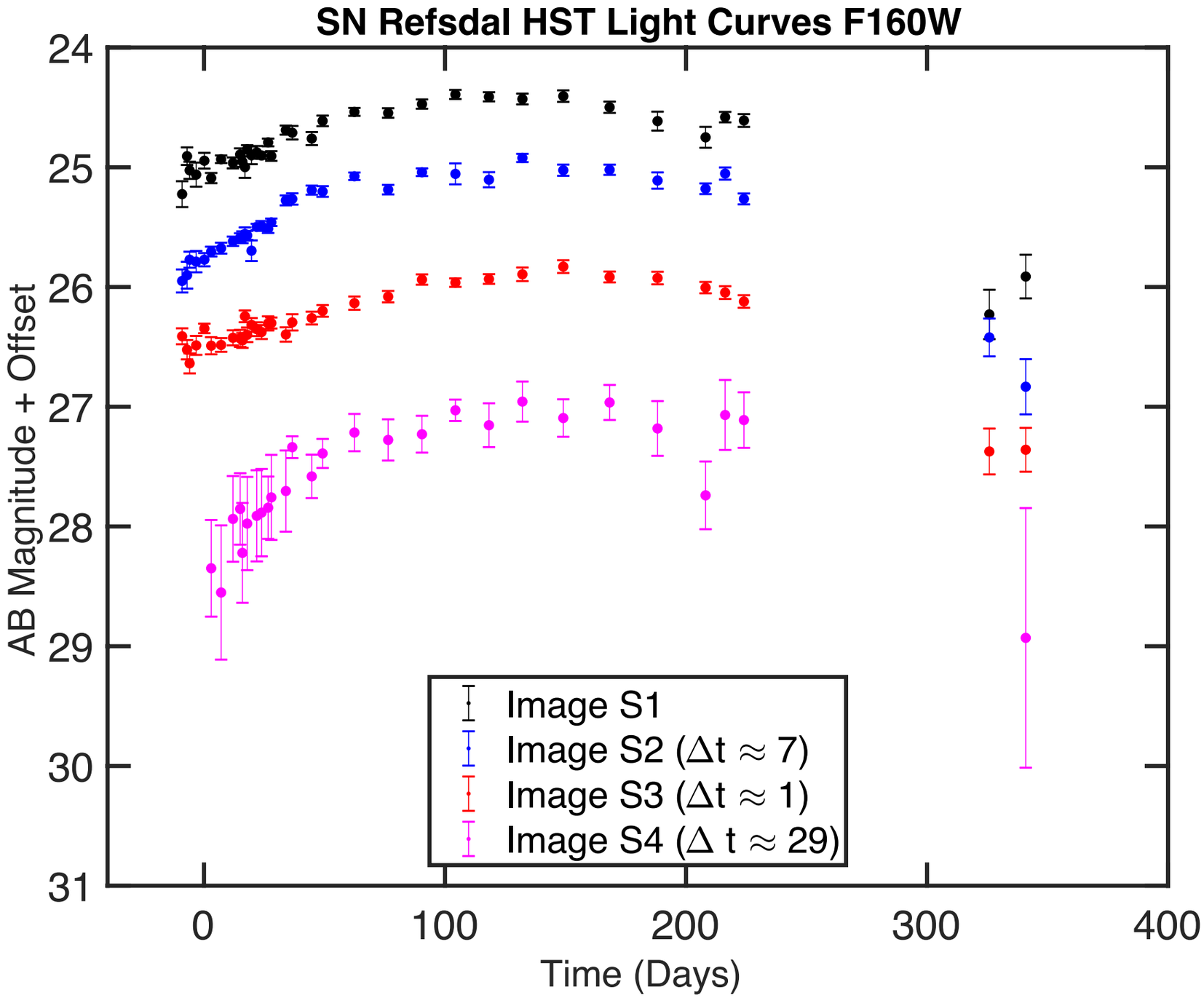
Time Series of SN brightnesses of each image: S1-S4

Hubble Space Telescope time series of SN Refsdal multiple images (Rodney et al. 2016)



Brightness Time Series [MJD = Modified Julian Day]

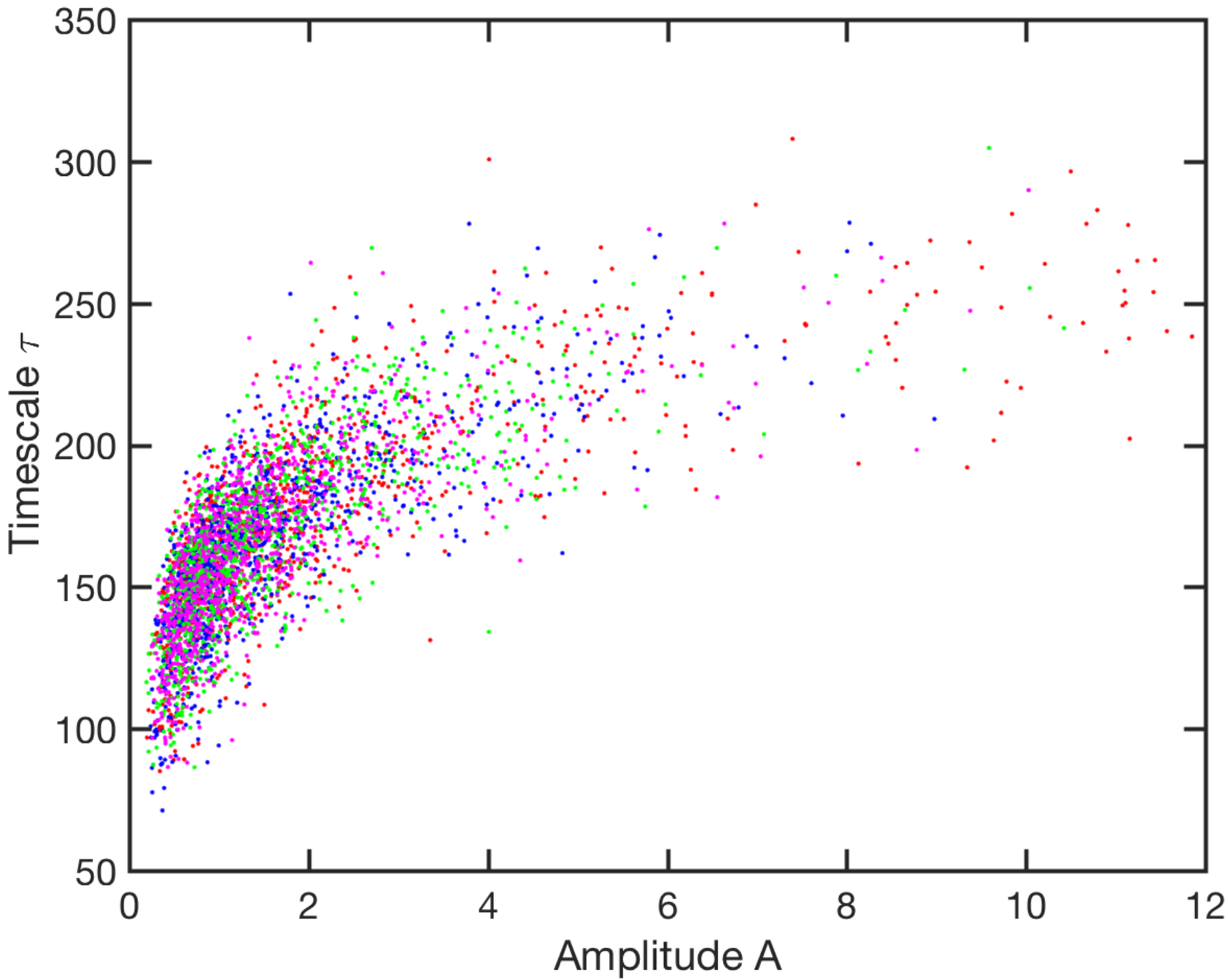
Hubble Space Telescope Time Series (light curves) of SN Refsdal at λ ≈ 1.6 μm



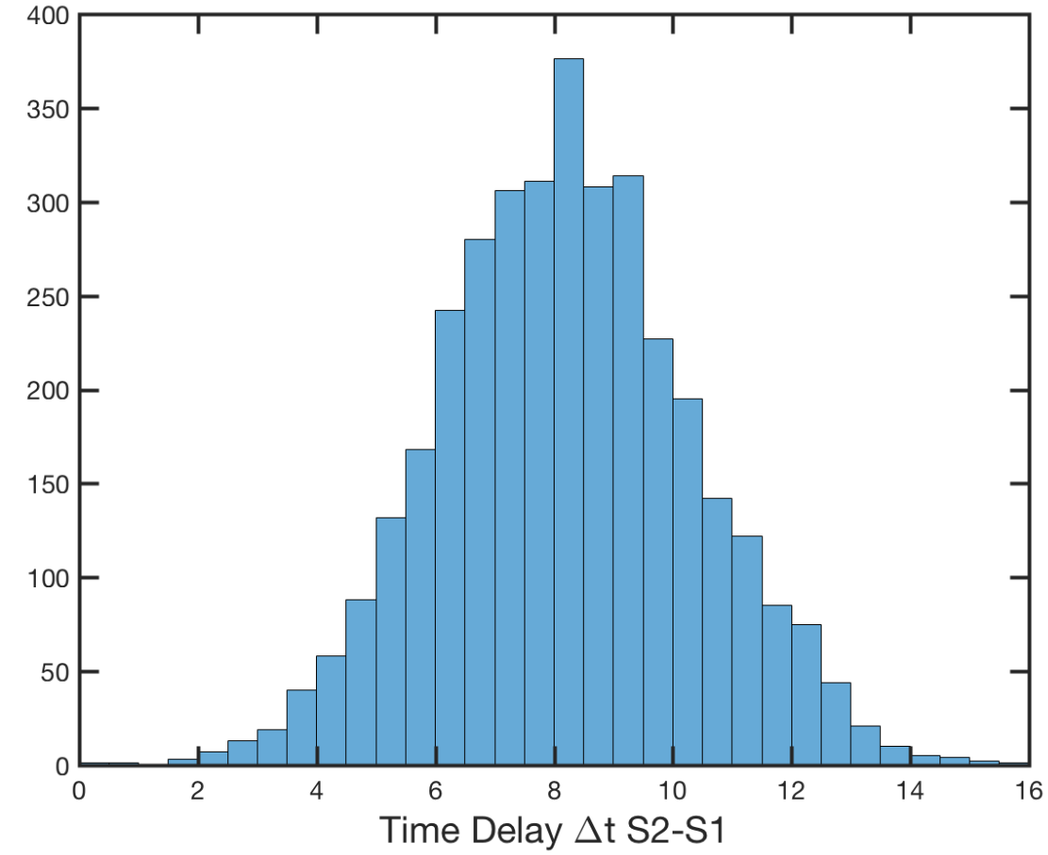
Rodney et al. 2016: Photometry & Time Delay Measurements of the first Einstein Cross Supernova

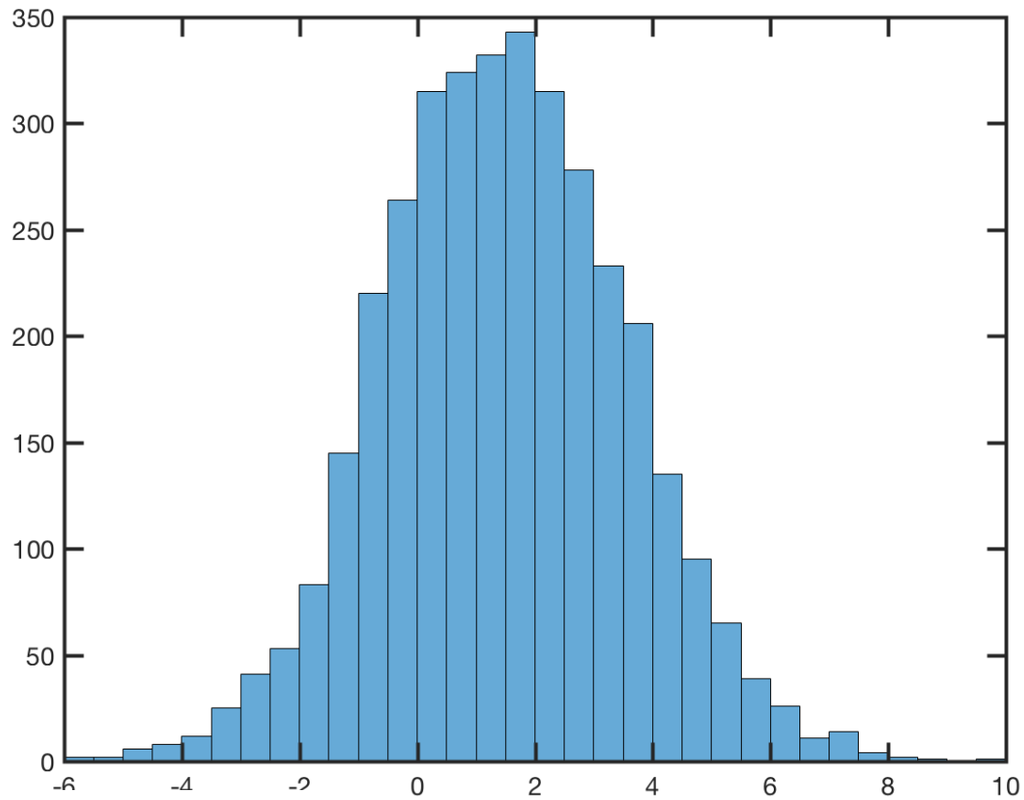
Write down Bayesian Model

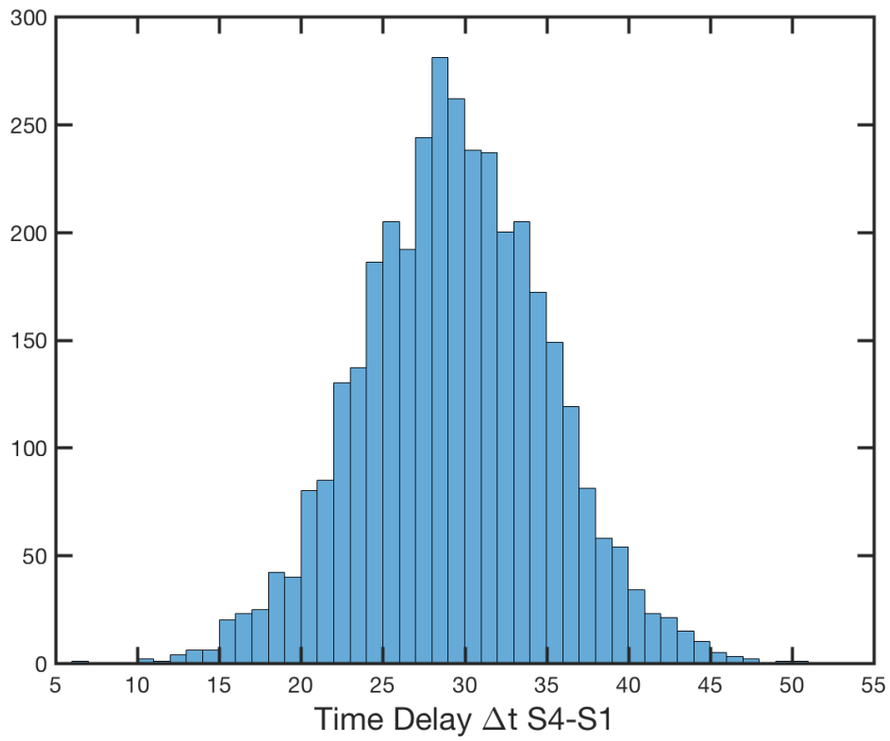
Metropolis-within-Gibbs results:
8 parameters, 4 chains
GP Hyperparameters



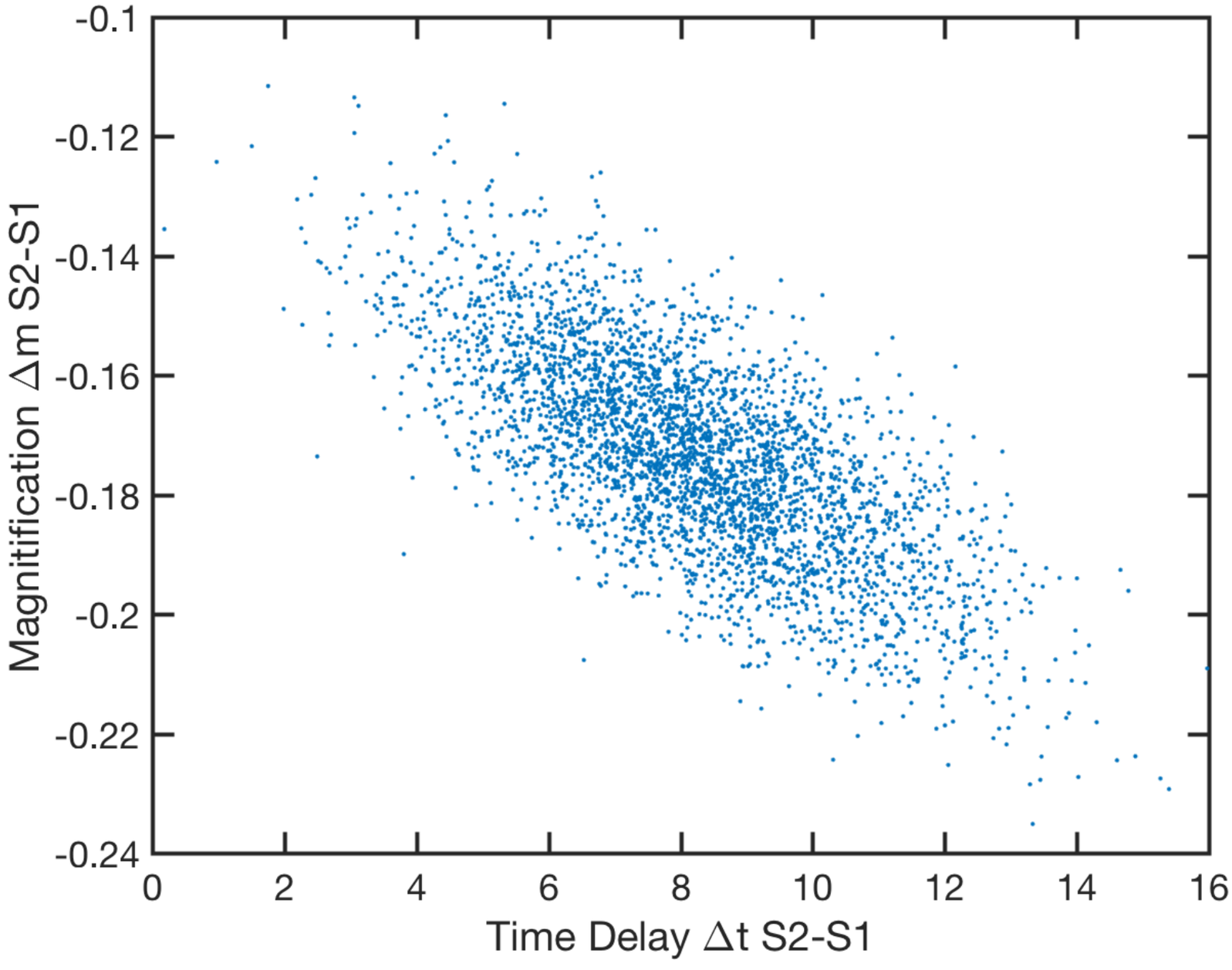
Metropolis-within-Gibbs results: 4 chains
Time Delays







Metropolis-within-Gibbs results: 4 chains
Time Delays vs Magnifications



Metropolis-within-Gibbs results: 4 chains
Deshifting the data