

Astrostatistics: 27 Feb 2019

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

- Example Class 2: Tue, 05 March, 12pm MR9
- Example Class 3: Tue, 12 March, 2pm MR 14, or?
 - Thu, 14 March, 12-2pm / 1-3pm / 3:30pm - 5:30 /?
- Example Class 4: Probably ~ April 29
- Some Lecture Demo basic codes are online in lecture_codes/
 - gibbs_example, metropolis1, metropolis2, metropgibbs_example
- Today: Finish sketch of MCMC Theory
 - See Robert & Casella “Monte Carlo Statistical Methods” for more details
- Hopefully! Begin Gaussian Processes in Astrophysics
- Case Study: application to gravitationally lensed quasars & supernovae

- Some remarks on mixed samplers and blocking
- Sketch of MCMC Theory

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
- 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

Review: Properties of Multivariate Gaussians

Full probability density: Σ is positive definite

$$N(\mathbf{f}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = [\det(2\pi\boldsymbol{\Sigma})]^{-1/2} \exp[-\frac{1}{2} (\mathbf{f} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{f} - \boldsymbol{\mu})]$$

Joint distribution of components:

$$\mathbf{f} = \begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix} \sim N \left(\begin{bmatrix} \mathbf{U}_0 \\ \mathbf{V}_0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_U & \boldsymbol{\Sigma}_{UV} \\ \boldsymbol{\Sigma}_{VU} & \boldsymbol{\Sigma}_V \end{bmatrix} \right)$$

If you observe/know/condition on V:

Conditional dist'n:

$$\mathbf{U} | \mathbf{V} \sim N(\mathbb{E}[\mathbf{U} | \mathbf{V}], \text{Var}[\mathbf{U} | \mathbf{V}])$$

Conditional Mean:

$$\mathbb{E}[\mathbf{U} | \mathbf{V}] = \mathbf{U}_0 + \boldsymbol{\Sigma}_{UV} \boldsymbol{\Sigma}_V^{-1} (\mathbf{V} - \mathbf{V}_0)$$

Conditional Variance:

$$\text{Var}[\mathbf{U} | \mathbf{V}] = \boldsymbol{\Sigma}_U - \boldsymbol{\Sigma}_{UV} \boldsymbol{\Sigma}_V^{-1} \boldsymbol{\Sigma}_{VU}$$

If \mathbf{V} = observed data, \mathbf{U} = unobserved parameters, then $P(\mathbf{U} | \mathbf{V})$ is a posterior pdf!

What is a Gaussian Process?

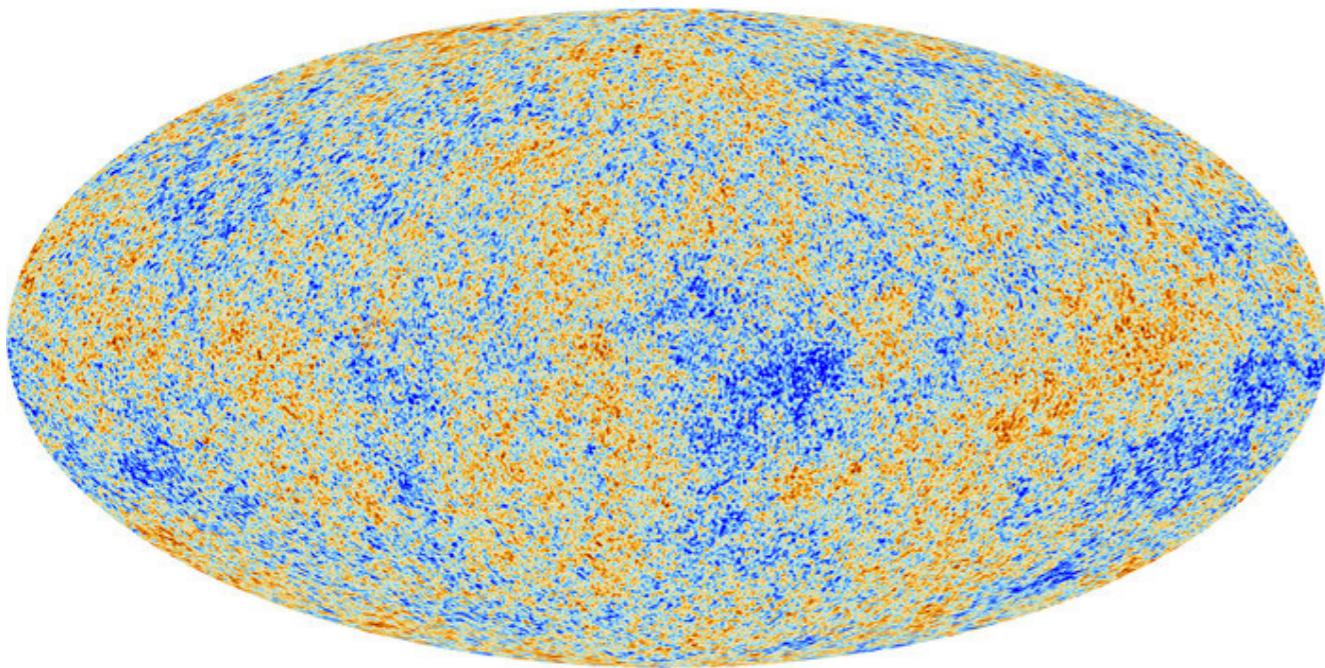
- A GP is a collection of random variables $\{f_t\}$, (typically with some ordering in time, space or wavelength), such that any finite subset of r.v.s have a jointly multivariate Gaussian distribution.
- Any vector $\mathbf{f} = \{f_t : t = 1 \dots N\}$ of a finite subset is multivariate Gaussian, therefore it is completely described by a mean $\mathbf{E}[\mathbf{f}]$ and covariance matrix $\mathbf{Var}[\mathbf{f}] = \mathbf{Cov}[\mathbf{f}, \mathbf{f}^T]$.
- Elements of the **covariance matrix** are determined by a function of the coordinates, e.g. $\text{Cov}[f_t, f_{t'}] = k(t, t')$, called the *covariance function* or *kernel*
- A Gaussian Process with mean function $m(t)$ is denoted:

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

What are GPs used for in astrophysics?

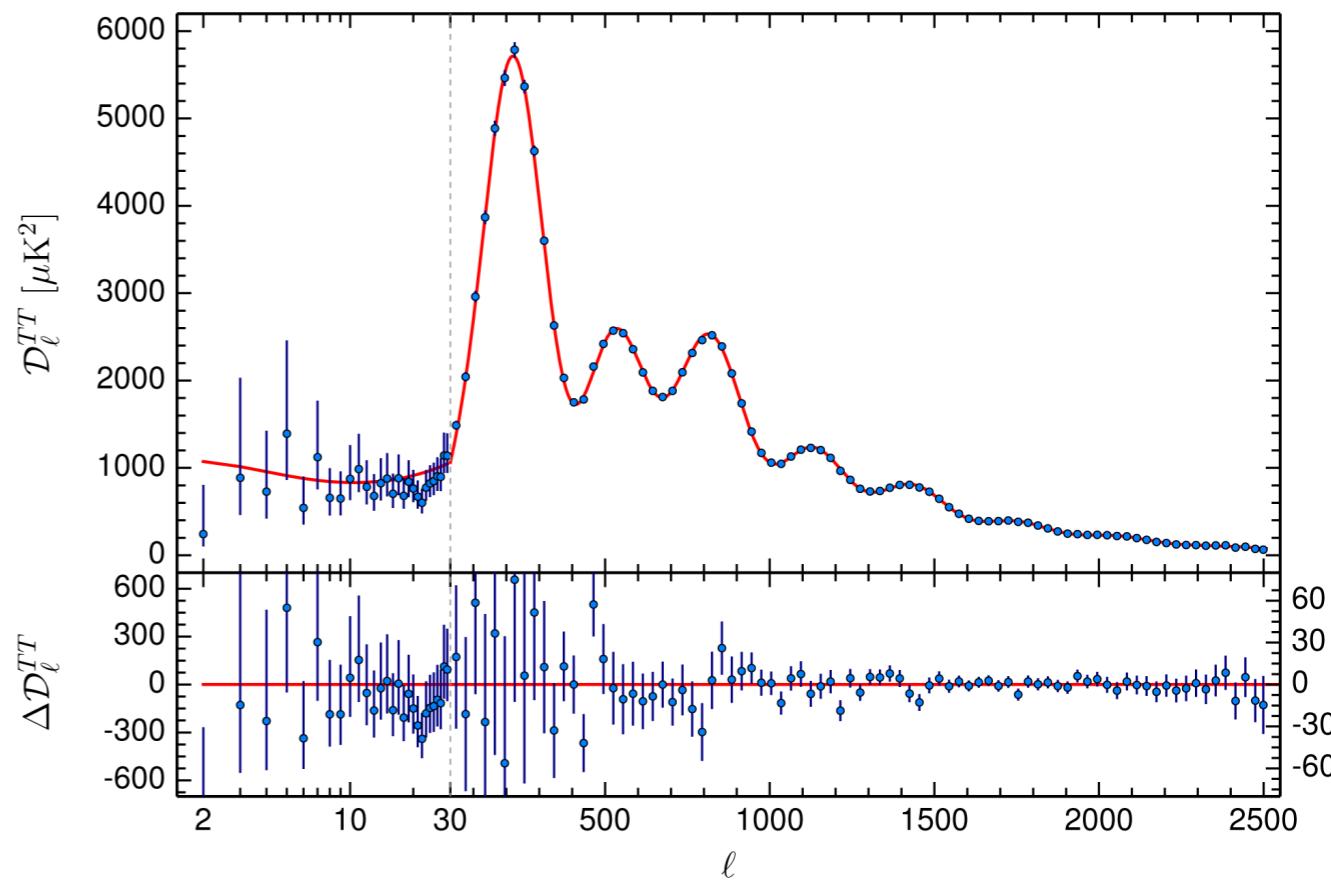
- Some physical models are actually very nearly Gaussian Processes (e.g. Cosmic Microwave Background is a GP on the sphere) or approximately so (damped random walks for quasar light curves)
- “Nonparametric” models: flexible functions to use when an accurate astrophysical function is not available or is imperfect
- Nonparametric \sim number of parameters grows with the dataset
- Interpolation/Emulation: To generate a smooth curve going through some observation or simulation points
- Correlated noise/error model: When you marginalise out the latent error function, you are effectively accounting for correlated noise over time/space/wavelength.
e.g. *D. Jones et al.* “Improving Exoplanet Detection Power: Multivariate Gaussian Process Models for Stellar Activity.” arXiv:1711.01318

Example: A Gaussian Process for the Spatial Variations of Intensity/Temperature



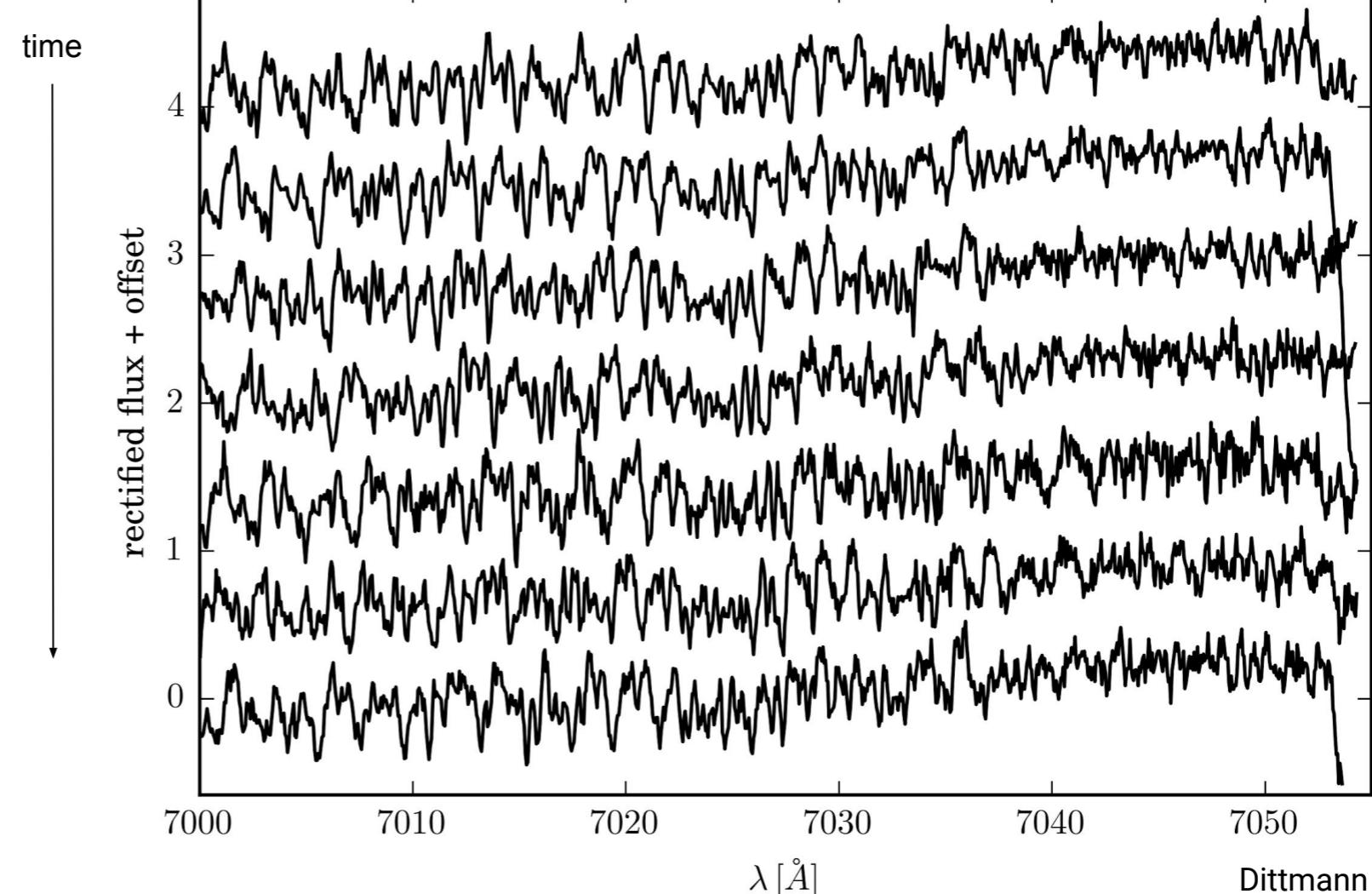
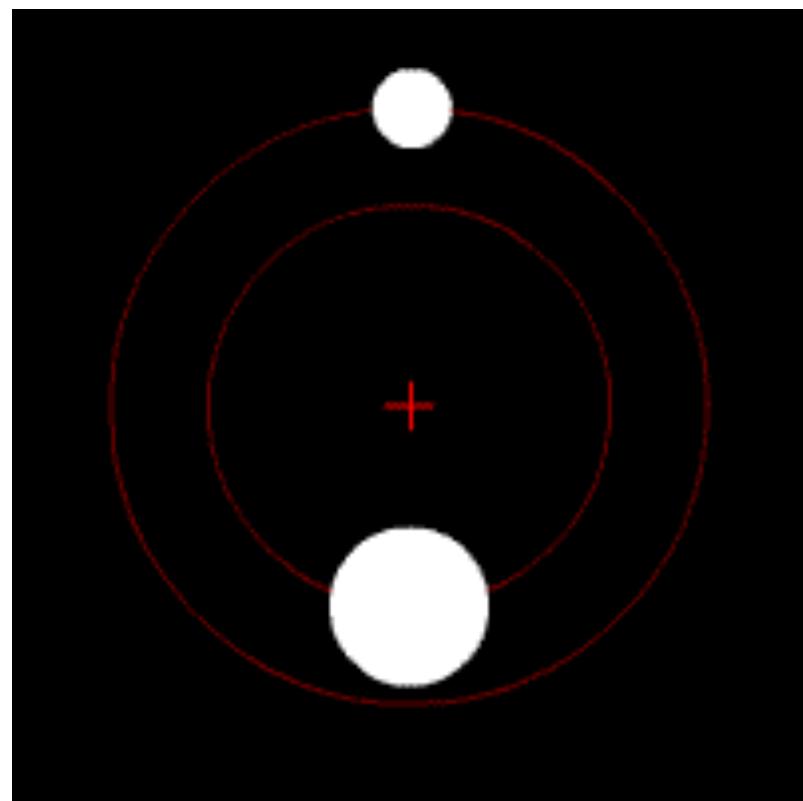
Cosmic Microwave
Background (Planck)
~ Gaussian Random Field
(mean = 2.7 K,
std dev $\sim 10^{-5}$)

Power Spectrum
(~Fourier Transform of
Covariance Function)



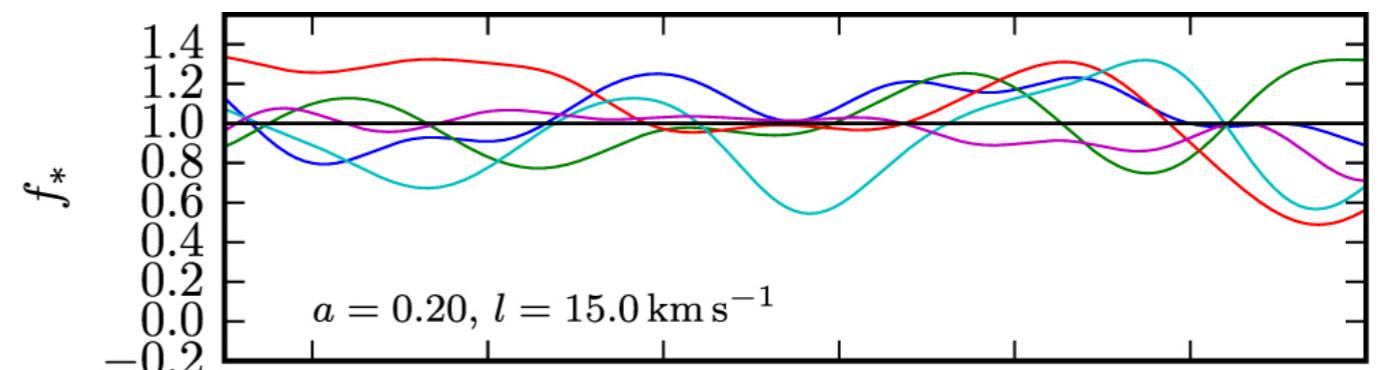
Example: *Astrostatistics Case Studies:*
Disentangling Time Series Spectra with Gaussian
Processes: Applications to Radial Velocity Analysis
(Czekala et al. 2017, ApJ, 840, 49. arXiv:1702.05652)

Raw Observations of the LP661-13 M4 Binary

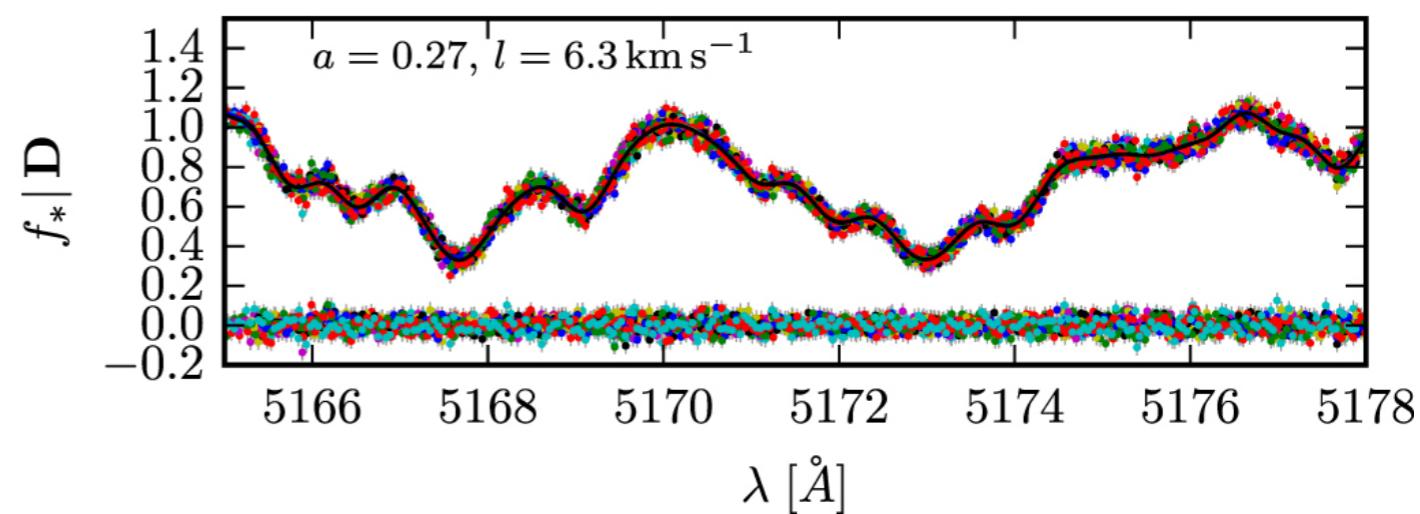
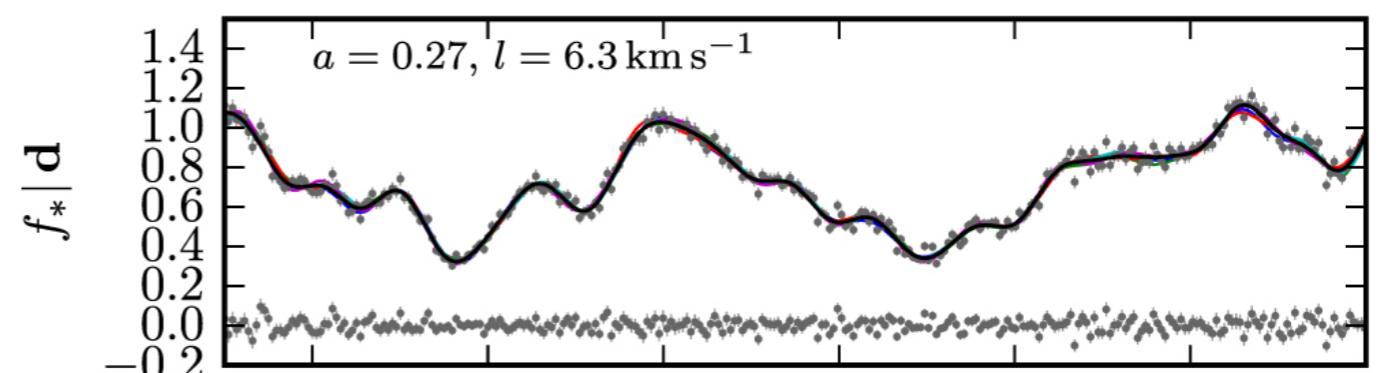
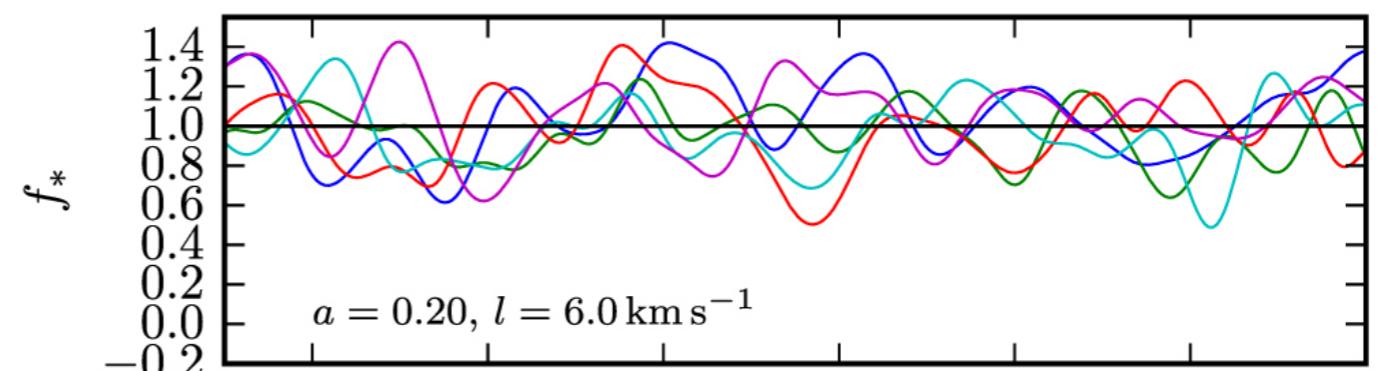


Example: Gaussian Process: Priors & Posteriors

GP prior (long/short correlation length scales)



GP Posterior
(conditioned on data spectrum \mathbf{d})
Inference of latent spectrum



Astrostatistics Case Study:

Bayesian Estimates of Astronomical Time

Delays Between Gravitationally Lensed Stochastic Light Curves

(Tak, Mandel et al. 2017, Annals of Applied Statistics, arXiv:1602.01462)

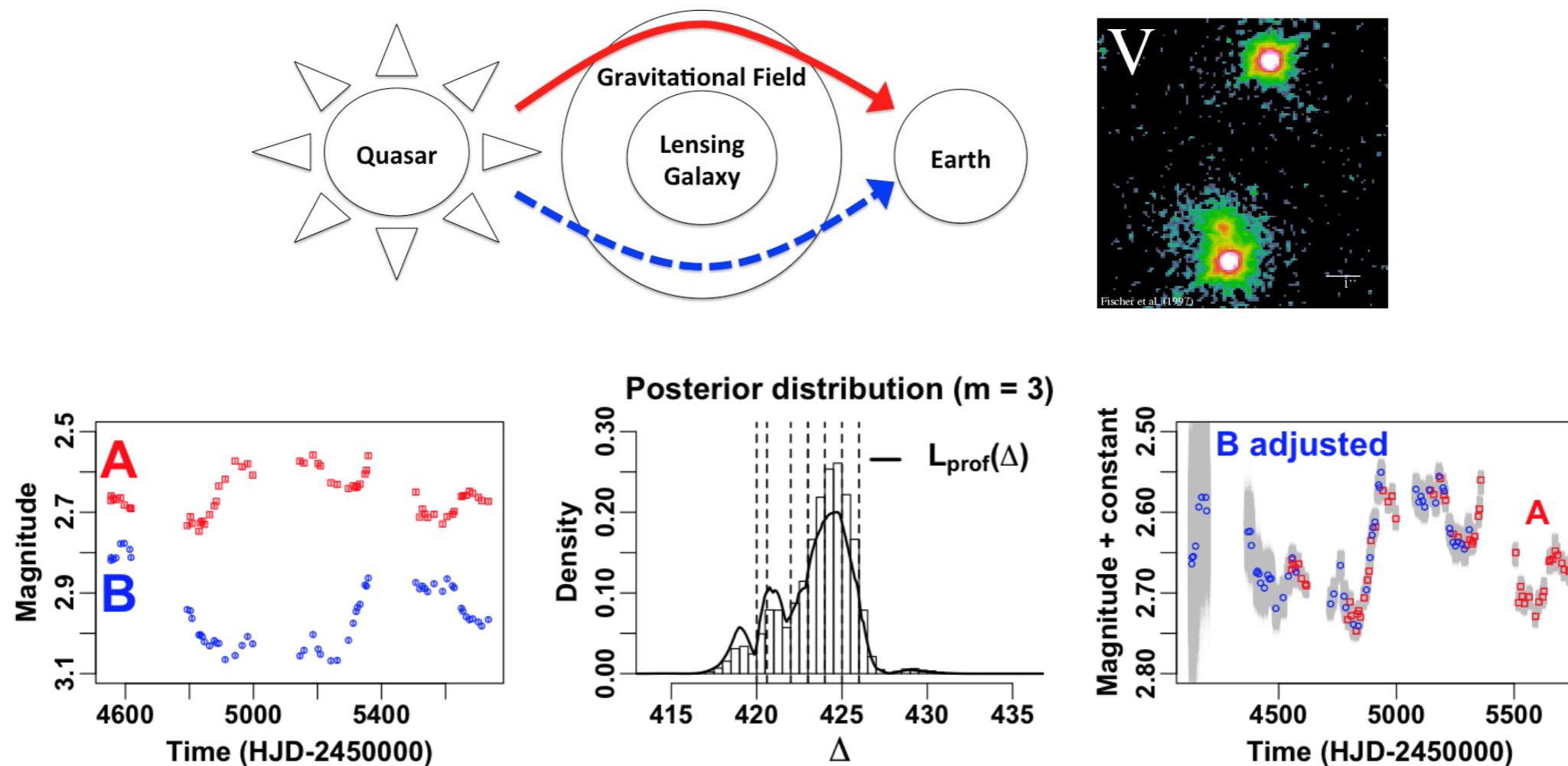
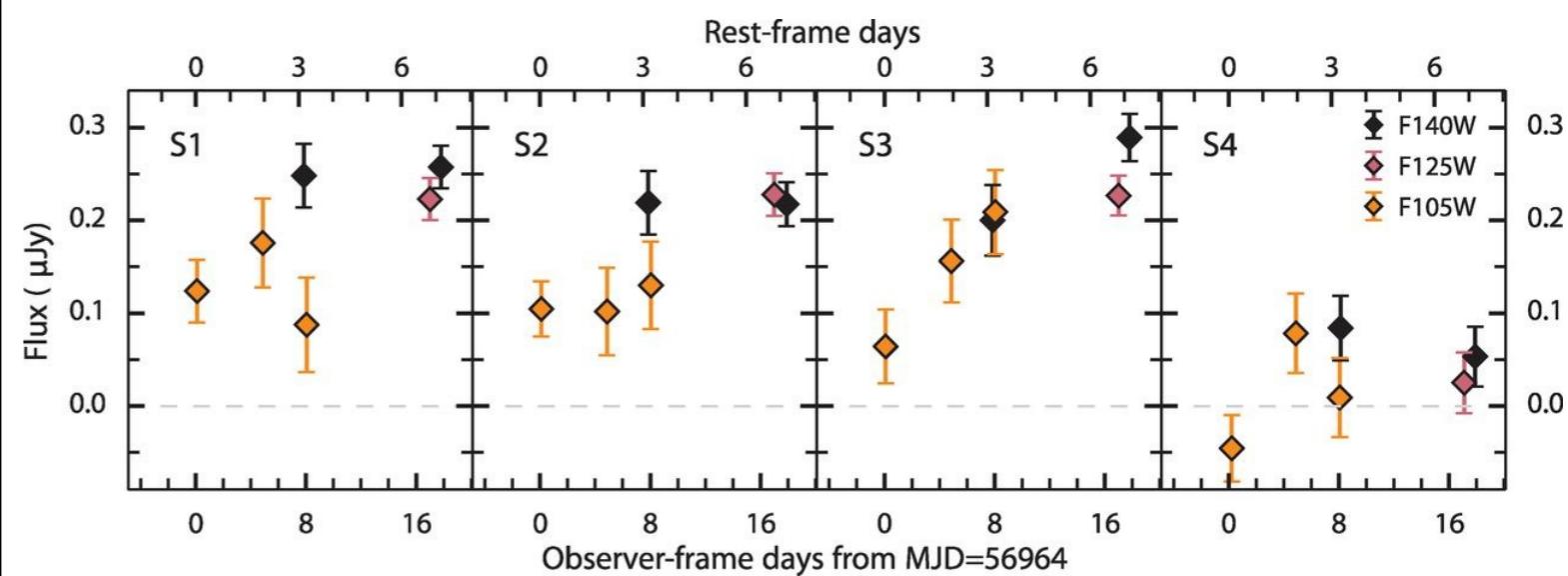
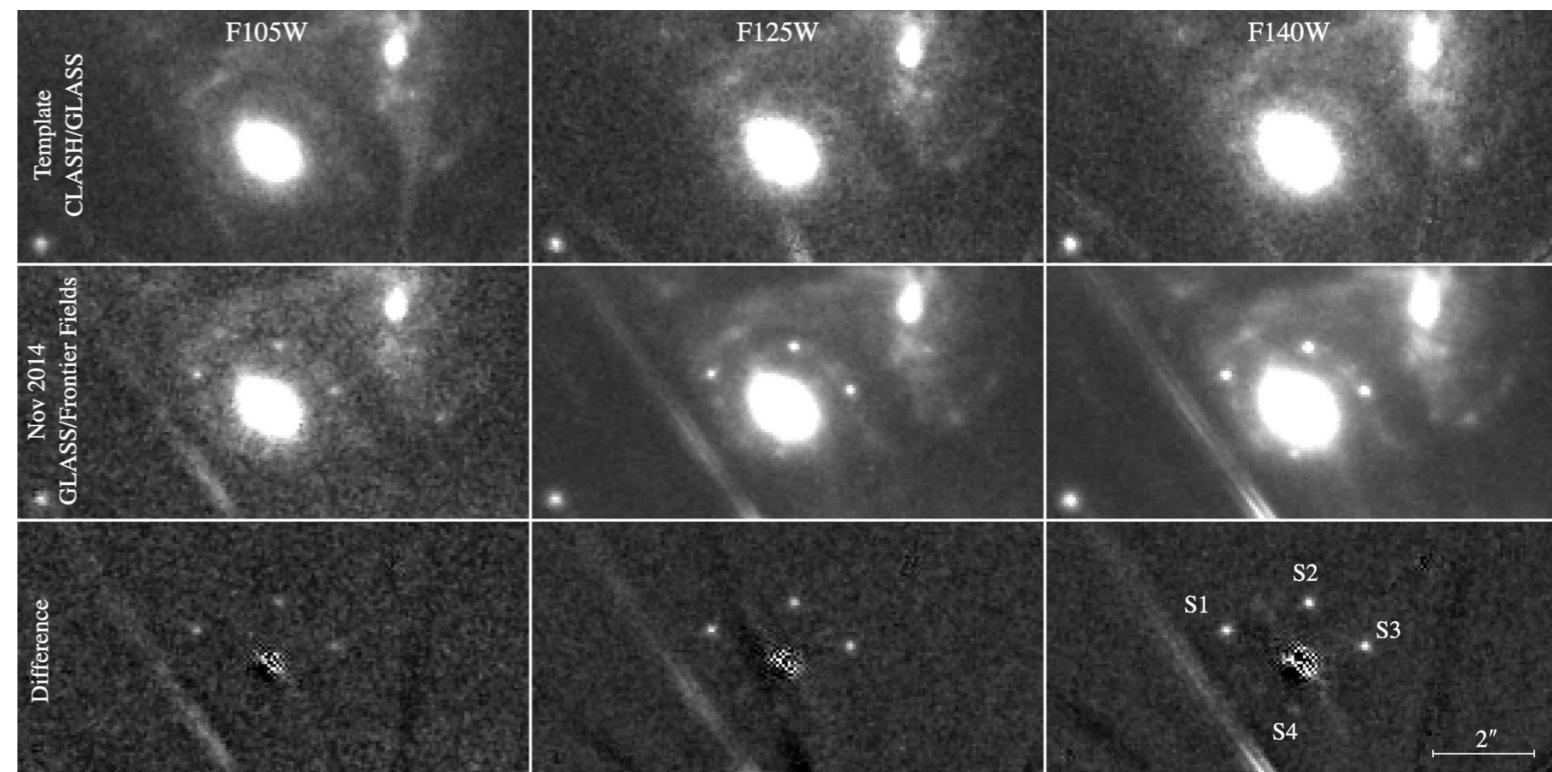
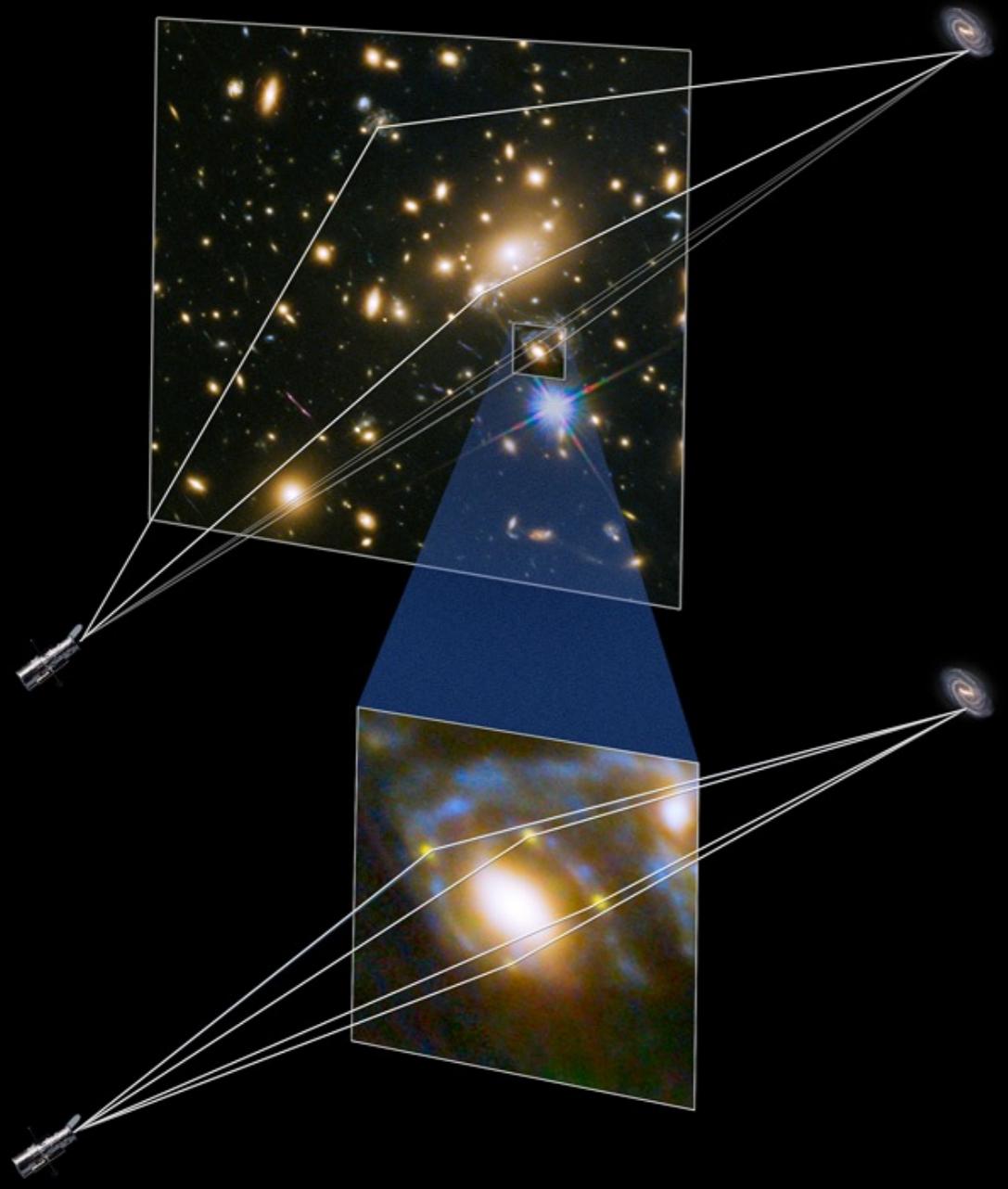


FIG 13. Observations of Quasar Q0957+561 from [Hainline et al. \(2012\)](#) are plotted in the first panel. The second panel exhibits the marginal posterior distribution of Δ with

Model the underlying latent light curve as a damped random walk (Ornstein-Uhlenbeck Process) to estimate time delays between time series to determine expansion rate of Universe (H_0)

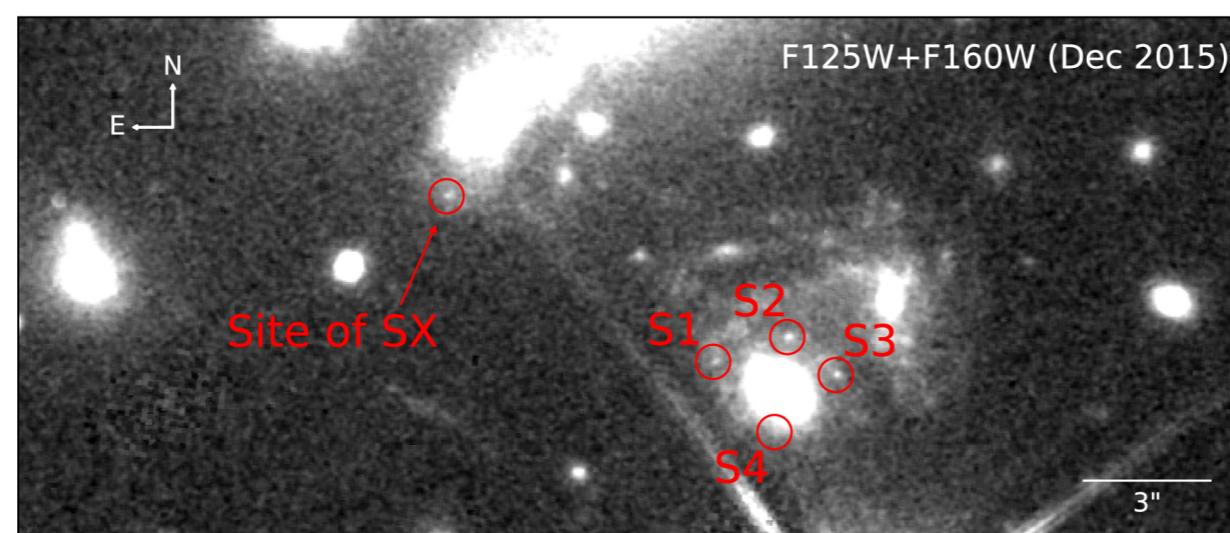
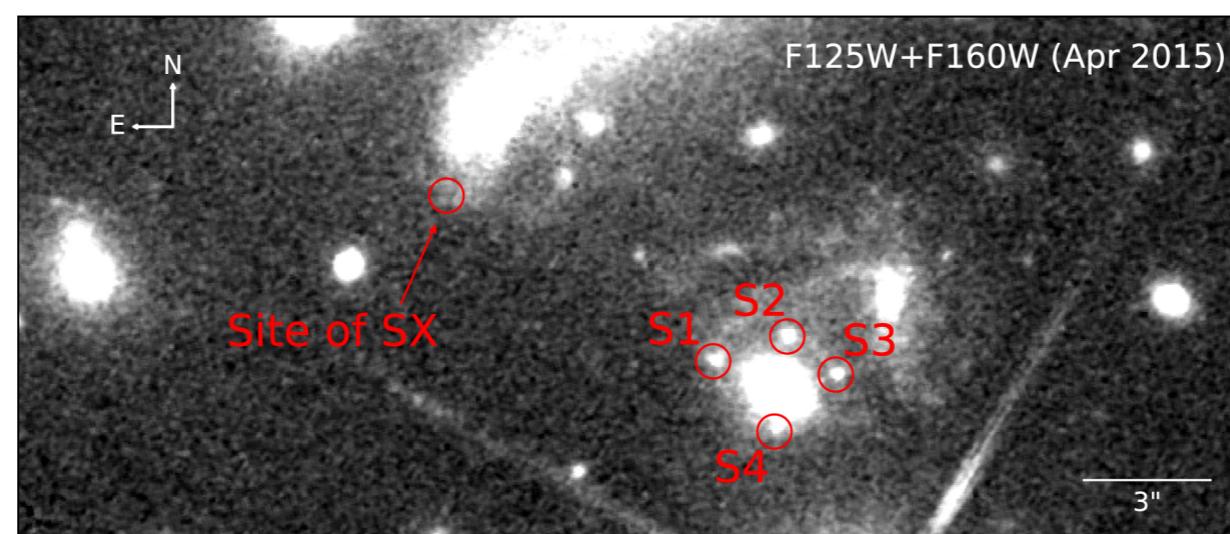
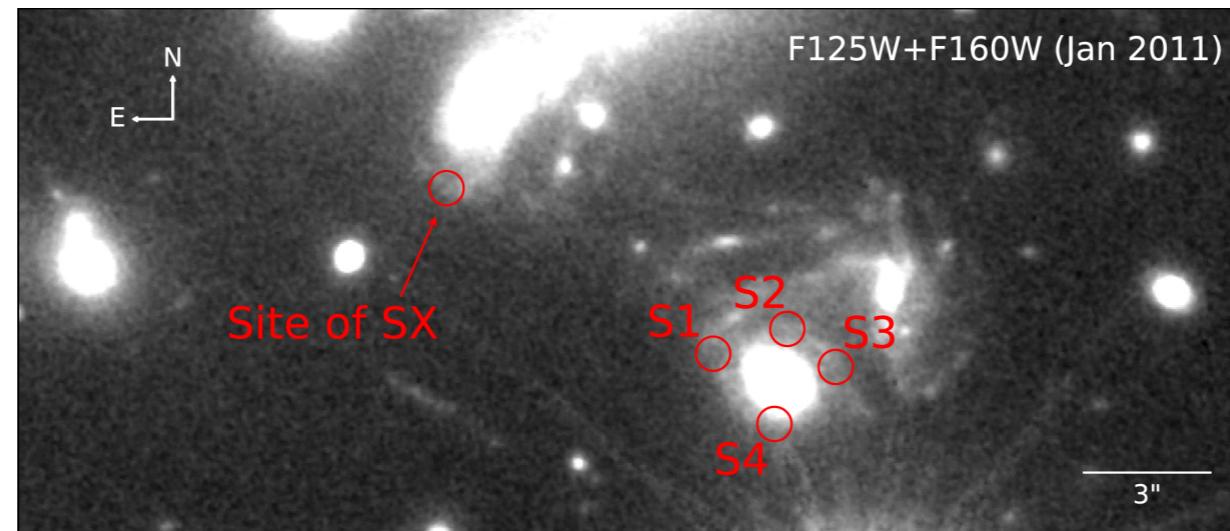
Today's Example: SN Refsdal

Hubble Sees Distant Supernova
Multiply Imaged by Foreground Galaxy Cluster

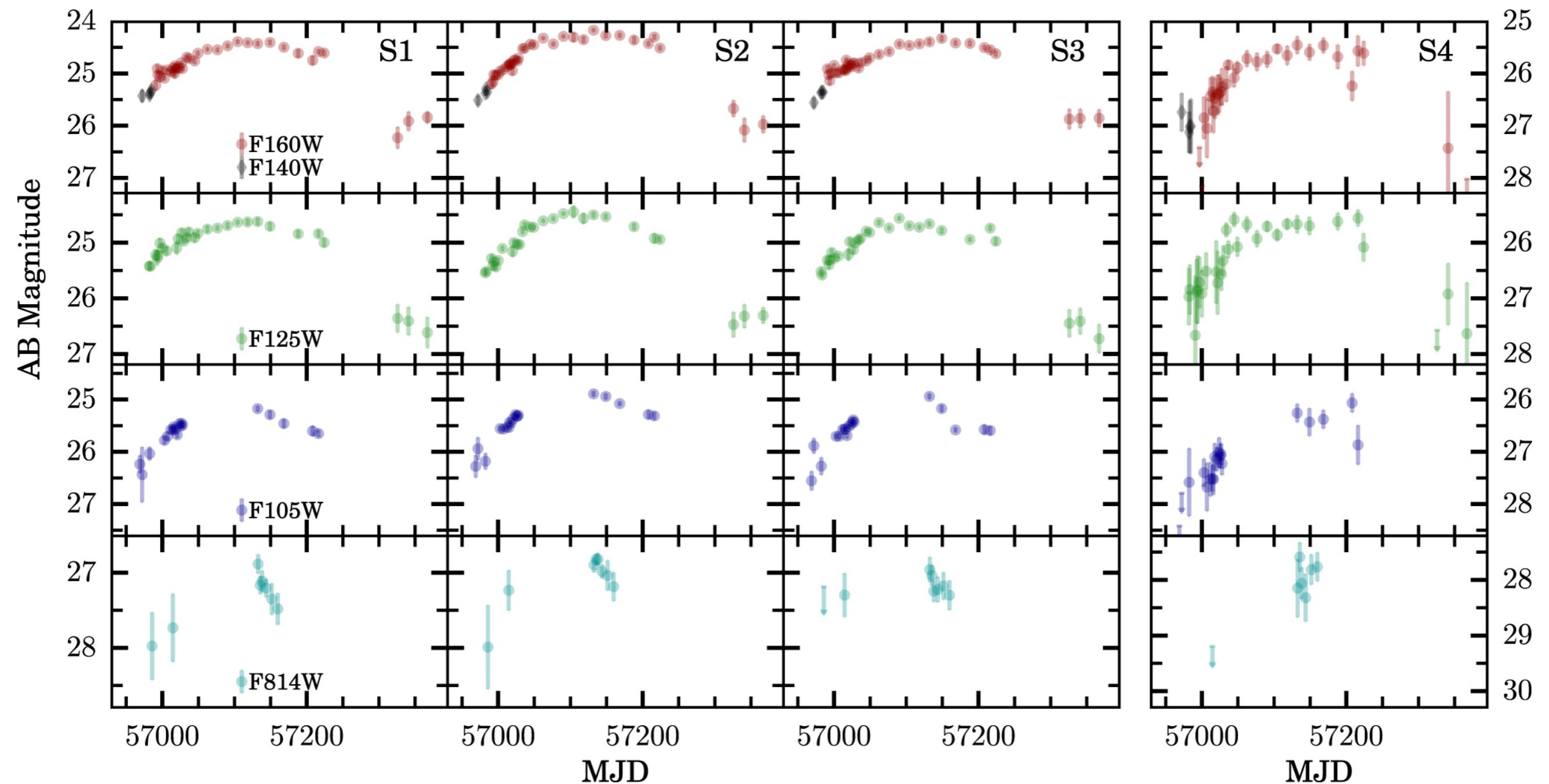


Time Series of SN brightnesses of each image: S1-S4

Prediction and Confirmation of the Reappearance of 5th Lensed Image of SN Refsdal (Kelly et al. 2015)

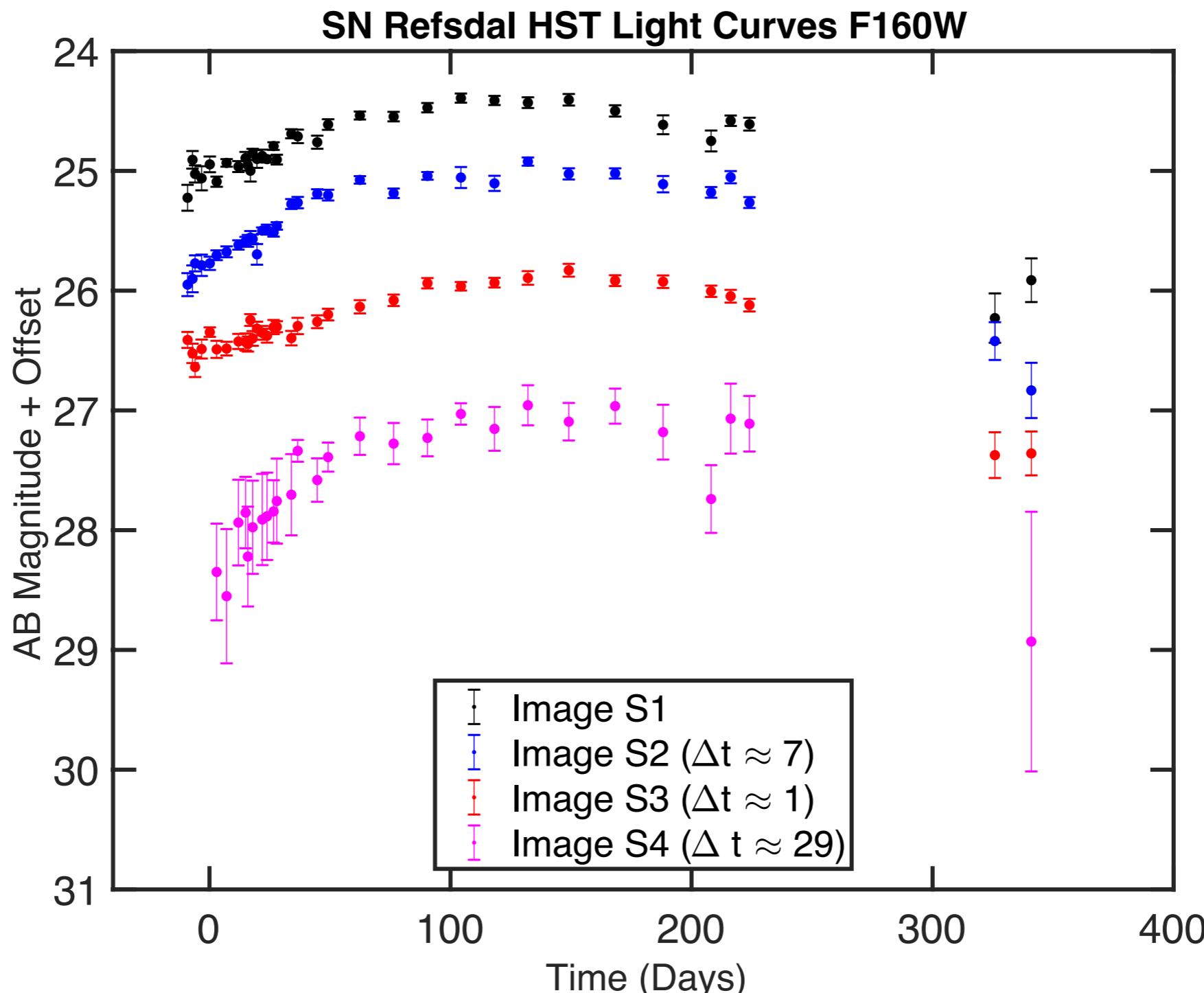


Hubble Space Telescope time series of SN Refsdal multiple images (Rodney et al. 2016)



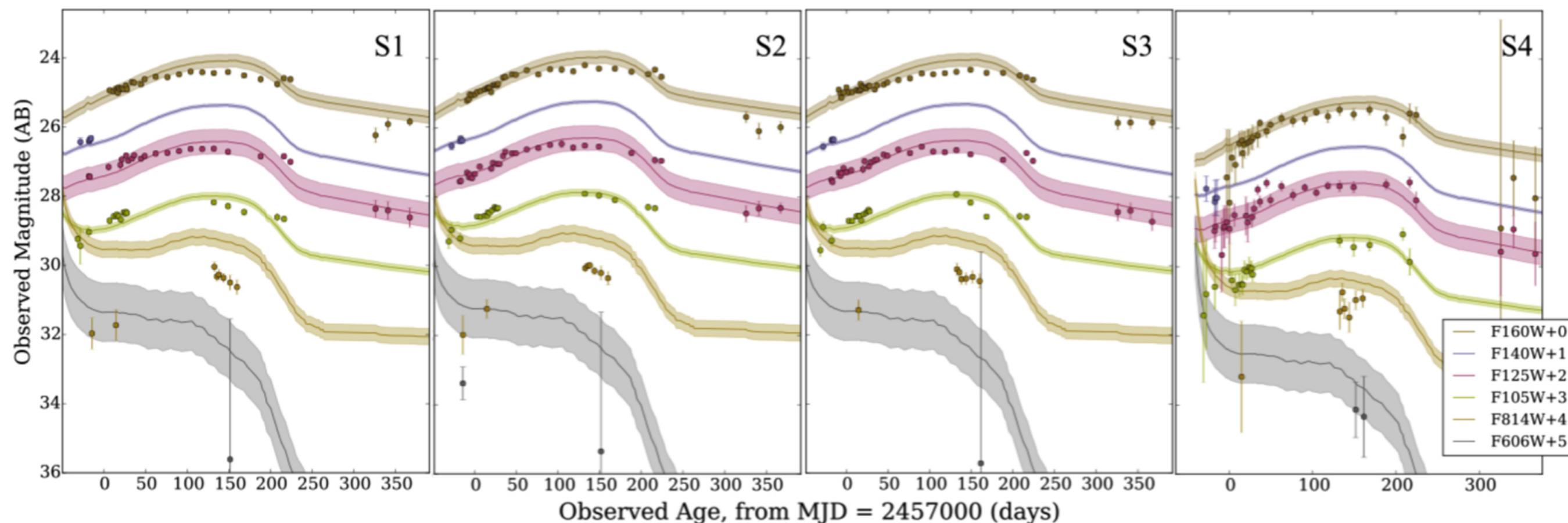
Brightness Time Series [MJD = Modified Julian Day]

Hubble Space Telescope Time Series (light curves) of SN Refsdal at $\lambda \approx 1.6 \mu\text{m}$



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

Doesn't fit well to a well-known SN light curve (SN 1987A)

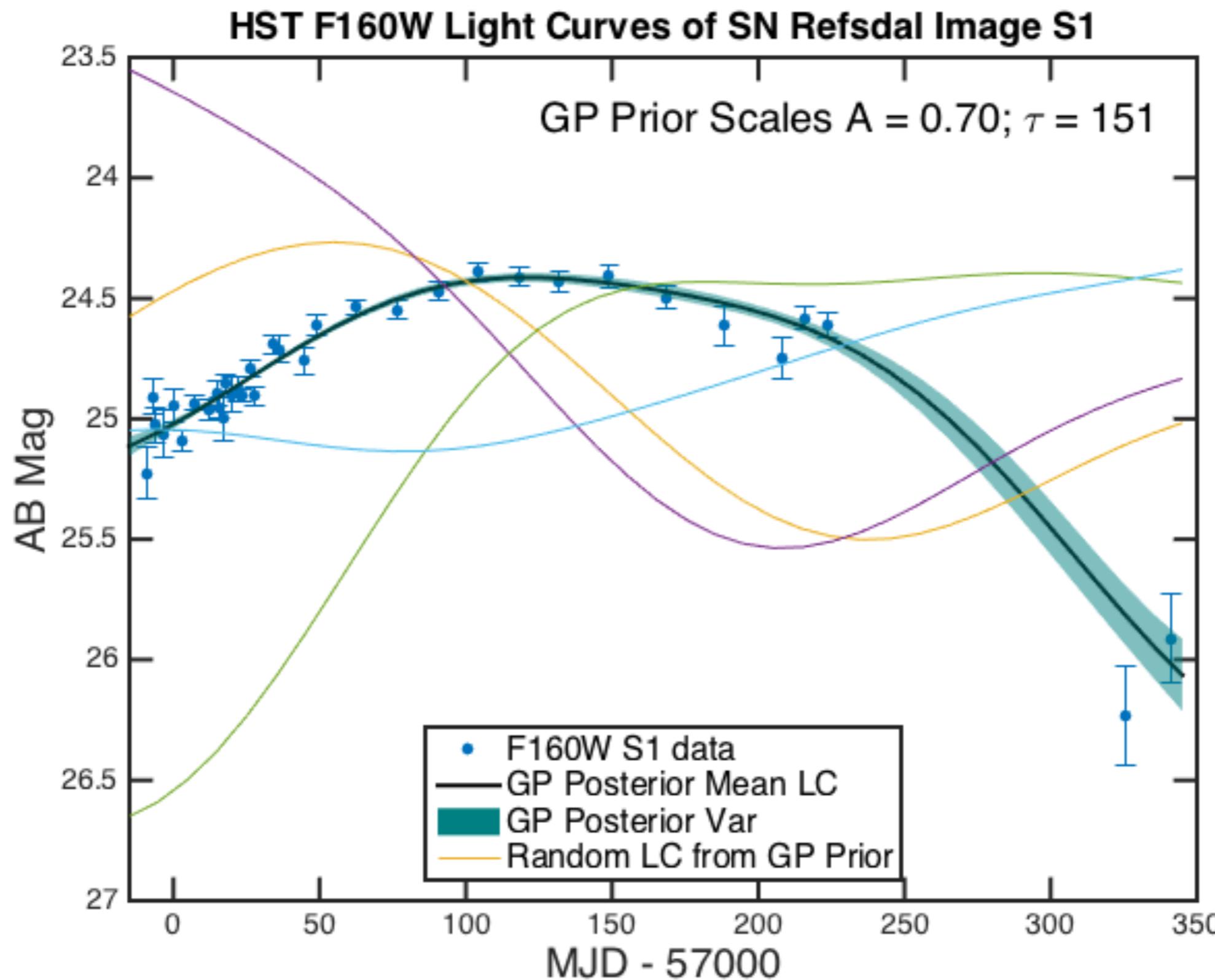


Our Strategy:

Model the underlying light curves as time delayed copies of one realisation of a smooth GP.

Use Bayesian inference to infer the time delays and latent LC simultaneously.

GP fit to a single time series



Gaussian Process as a prior on functions

A grid of times: $\mathbf{t} = (t^1, \dots, t^i, \dots, t^N)^T$

A vector of function values on the time grid:

$$\mathbf{f} = (f(t^1), \dots, f(t^i), \dots, f(t^N))^T$$

Assume a Squared Exponential kernel / covariance function

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

Assume a constant prior mean function:

$$\mathbb{E}[f(t)] = m(t) = c = 25.5 \quad (\text{Often assume zero-mean } c = 0)$$

Gaussian Process as a prior on functions

Prior on function: $P(\mathbf{f} | A, \tau) = N(\mathbf{f} | \mathbf{1}_c, \mathbf{K})$

Drawing from Prior: $\mathbf{f} | A, \tau \sim N(\mathbf{1}_c, \mathbf{K})$

Covariance Matrix \mathbf{K} populated by evaluating the kernel:

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

For all pairs of points in \mathbf{t} :

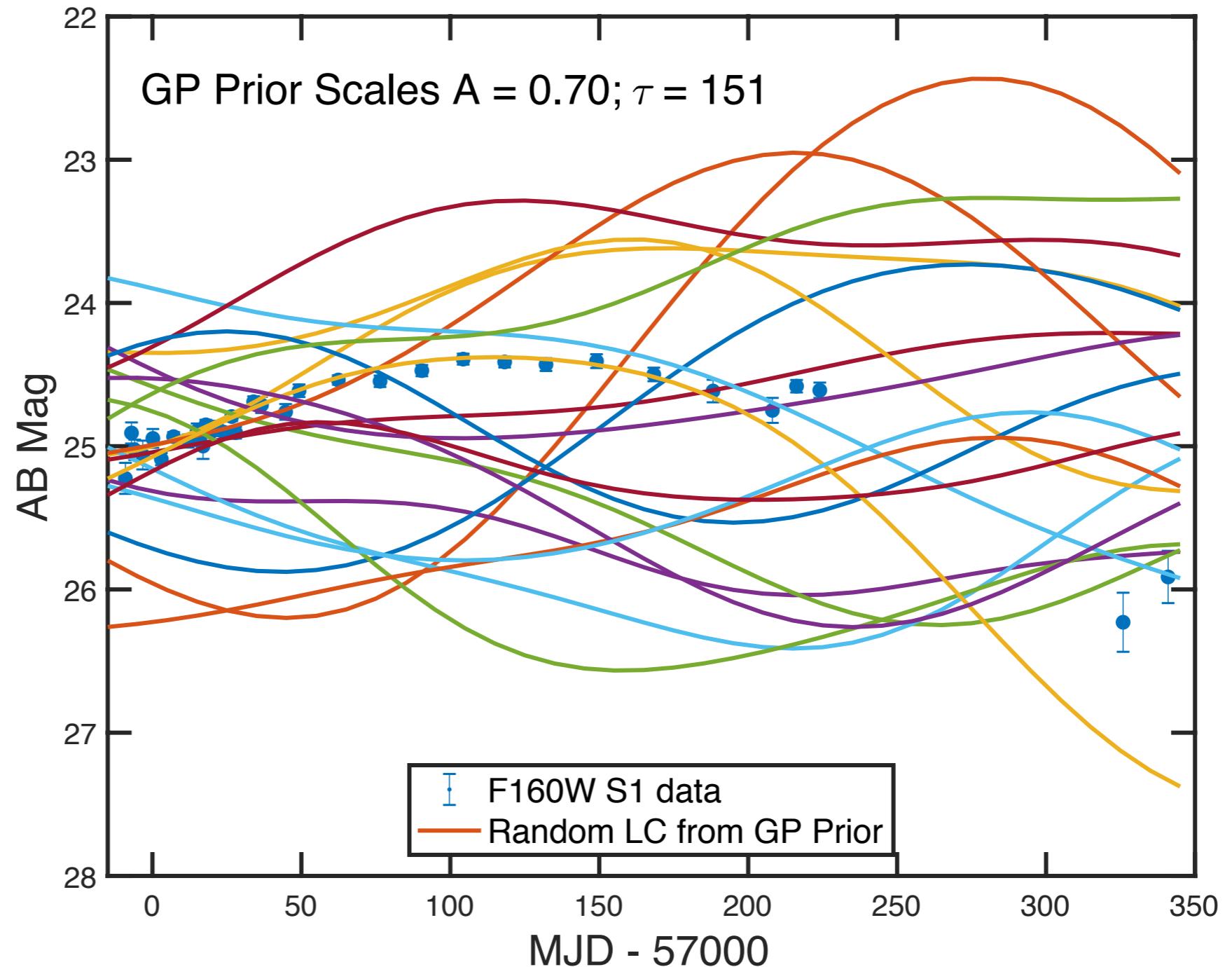
$$K_{ij} = k(t_i, t_j) = A \exp(-|t_i - t_j|^2 / \tau^2)$$

Drawing random functions from GP prior

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

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

Long
Characteristic
Timescale
 $\tau = 151$

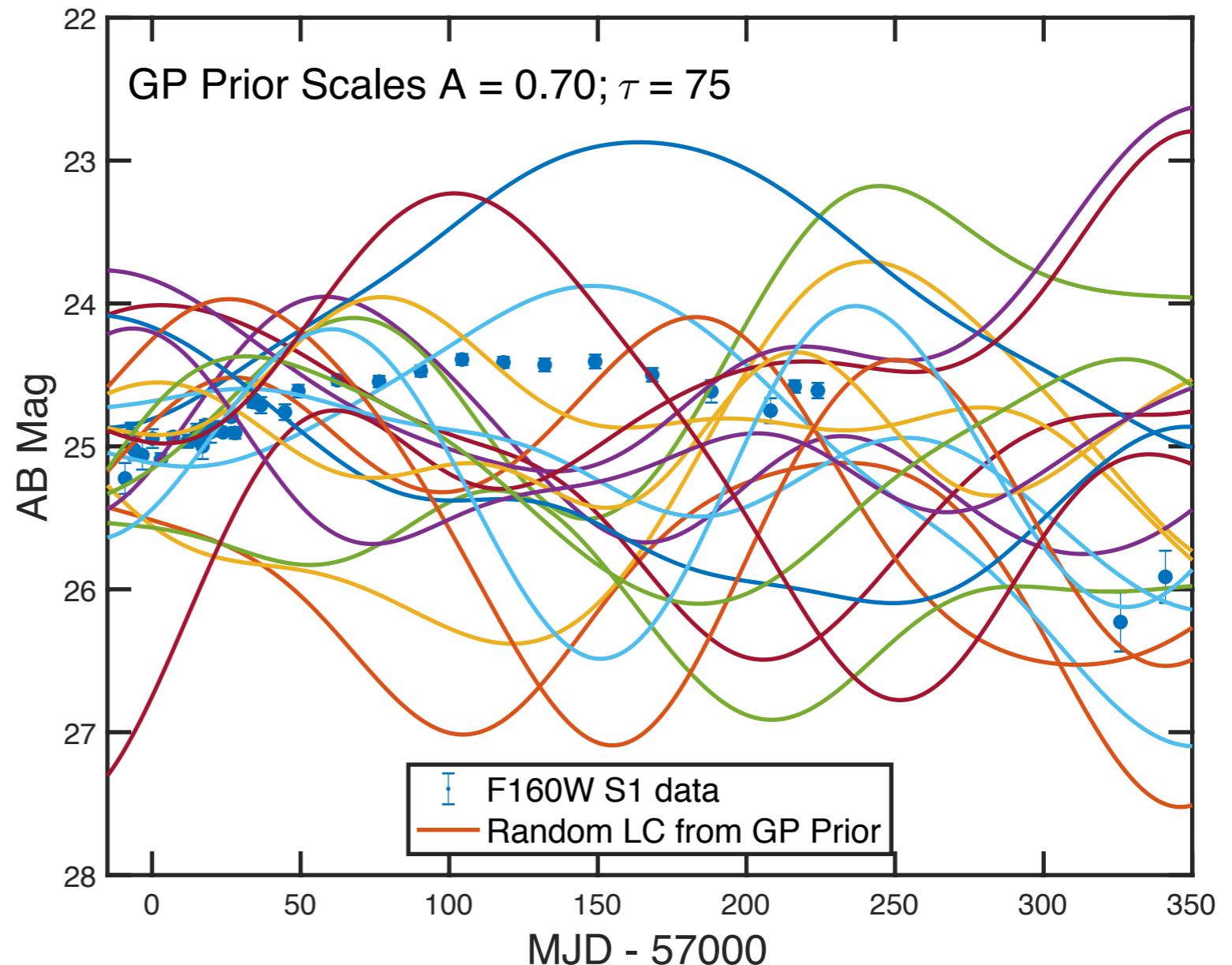


Drawing random functions from GP prior

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

$$f | A, \tau \sim N(1c, K)$$

Shorter
Characteristic
Timescale
 $\tau = 75$



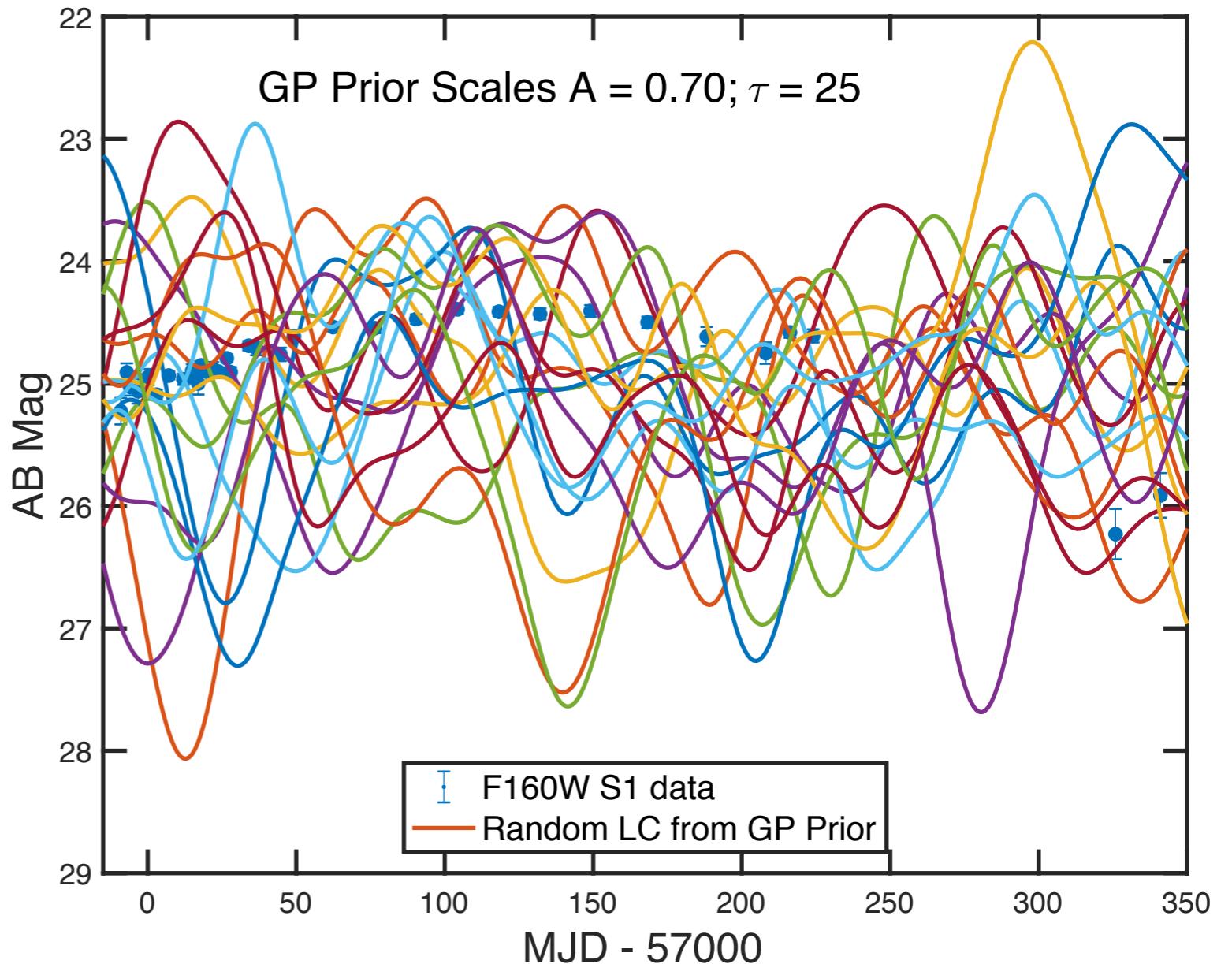
Drawing random functions from GP prior

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

$$f | A, \tau \sim N(1c, K)$$

Even Shorter
Characteristic
Timescale

$$\tau = 25$$



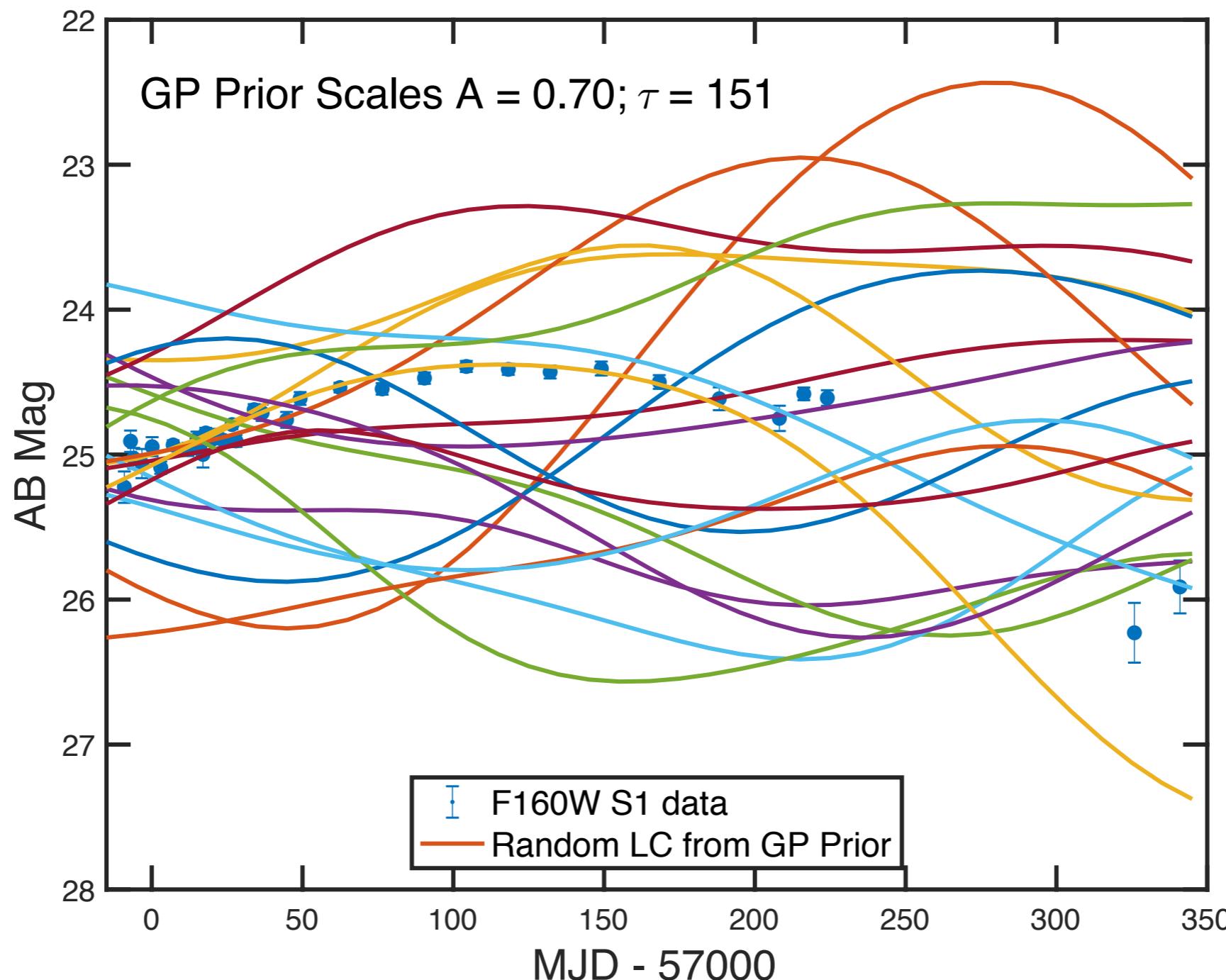
Fitting a GP to data

1. If we knew the characteristic scales of the kernel (A , τ^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?

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

$$f | A, \tau \sim N(1c, K)$$



Posterior Inference with GPs

Estimating the underlying curve:

f_o = observed points at times t_o (training set)

f_* = function at unobserved times t_* (prediction or test set)

Joint:

$$\begin{pmatrix} f_o \\ f_* \end{pmatrix} \sim N \left(\begin{bmatrix} 1c \\ 1c \end{bmatrix}, \begin{bmatrix} K(t_o, t_o) & K(t_*, t_o) \\ K(t_o, t_*) & K(t_*, t_*) \end{bmatrix} \right)$$

Populating the Covariance Matrix

$K(t, t')$ has i,j-th entry = $k(t_i, t'_j)$

Using the assumed kernel function

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

Posterior Inference with GPs

Estimating the underlying curve:

f_o = observed points at times t_o

f_* = function at unobserved times t_*

Jointly Gaussian: $\begin{pmatrix} f_o \\ f_* \end{pmatrix} \sim N \left(\begin{bmatrix} 1c \\ 1c \end{bmatrix}, \begin{bmatrix} K(t_o, t_o) & K(t_*, t_o) \\ K(t_o, t_*) & K(t_*, t_*) \end{bmatrix} \right)$

Posterior is also Gaussian $f_*|f_o \sim N(\mathbb{E}[f_*|f_o], \text{Var}[f_*|f_o])$

Posterior Mean:

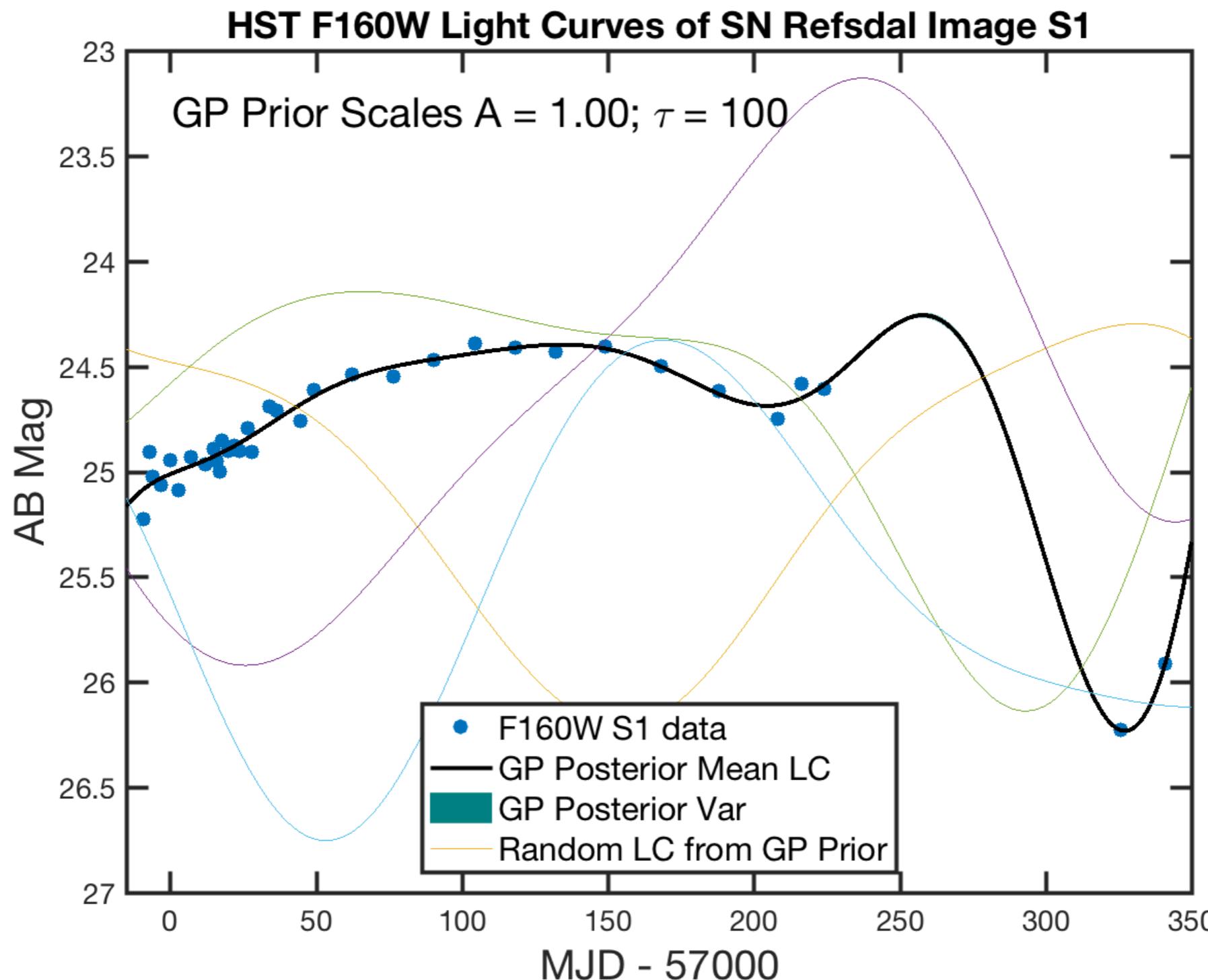
$$\mathbb{E}[f_*|f_o] = 1c + K(t_*, t_o)K(t_o, t_o)^{-1}(f_o - 1c)$$

Posterior Co(variance):

$$\text{Var}[f_*|f_o] = K(t_*, t_*) - K(t_*, t_o)K(t_o, t_o)^{-1}K(t_o, t_*)$$

Posterior Inference with GPs

Estimating the underlying curve:



Accounting for Measurement Error

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

\mathbf{y}_o are measured values of \mathbf{f}_o at times t_o

\mathbf{W} is measurement covariance matrix
(often diagonal for independent noise)

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

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

Now can calculate function prediction at unobserved points

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

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

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

f_o = function at observed times t_o

$$f_o \sim N[\mathbf{1}_c, \mathbf{K}(t_o, t_o)]$$

GP of Measurement Error

$$\mathbf{y}_o | f_o \sim N(f_o, \mathbf{W})$$

Same as: (mean-zero noise)

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

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: $f_o \sim N[\mathbf{1}_c, K(t_o, t_o)]$

Measurement Process: $y_o | f_o \sim N(f_o, W)$

$$y_o = f_o + \epsilon \quad \epsilon \sim N(\mathbf{0}, W)$$

$$\text{Cov}(y_o, y_o) = \text{Cov}(f_o, f_o) + \text{Cov}(\epsilon, \epsilon) + 2 \text{Cov}(f_o, \epsilon)$$

$$\text{Cov}(f_o, f_o) = K(t_o, t_o) \quad (\text{GP of intrinsic curve})$$

$$\text{Cov}(\epsilon, \epsilon) = W \quad (\text{measurement noise})$$

(the two processes are uncorrelated)

$$2 \text{Cov}[f_o, \epsilon] \equiv \text{Cov}[f_o, \epsilon][f_o, \epsilon] = 0$$

Therefore: $\text{Cov}[y_o, y_o] = K(t_o, t_o) + W$

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

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

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

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

Similar arguments for:

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

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

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

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

$$\begin{pmatrix} f_o \\ f_* \end{pmatrix} \sim N \left(\begin{bmatrix} 1_c \\ 1_c \end{bmatrix}, \begin{bmatrix} K(t_o, t_o) & K(t_*, t_o) \\ K(t_o, t_*) & K(t_*, t_*) \end{bmatrix} \right)$$

$$\begin{pmatrix} y_o \\ f_* \end{pmatrix} \mid \begin{pmatrix} f_o \\ f_* \end{pmatrix} \sim N \left(\begin{bmatrix} f_o \\ f_* \end{bmatrix}, \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix} \right)$$

$$P(V) \& P(U|V) \longrightarrow P(U)$$

$$\begin{pmatrix} y_o \\ f_* \end{pmatrix} \sim N \left(\begin{bmatrix} 1_c \\ 1_c \end{bmatrix}, \begin{bmatrix} K(t_o, t_o) + W & K(t_*, t_o) \\ K(t_o, t_*) & K(t_*, t_*) \end{bmatrix} \right)$$

Now can calculate function prediction at unobserved points

$$f_* | y_o \sim N(\mathbb{E}[f_* | y_o], \text{Var}[f_* | y_o])$$

Using conditional properties of Gaussian as before

Accounting for Measurement Error:

$$\begin{pmatrix} \mathbf{y}_o \\ f_* \end{pmatrix} \sim N \left(\begin{bmatrix} 1c \\ 1c \end{bmatrix}, \begin{bmatrix} \mathbf{K}(t_o, t_o) + \mathbf{W} & \mathbf{K}(t_*, t_o) \\ \mathbf{K}(t_o, t_*) & \mathbf{K}(t_*, t_*) \end{bmatrix} \right)$$

Now can calculate function prediction at unobserved points

$$f_* | \mathbf{y}_o \sim N(\mathbb{E}[f_* | \mathbf{y}_o], \text{Var}[f_* | \mathbf{y}_o])$$

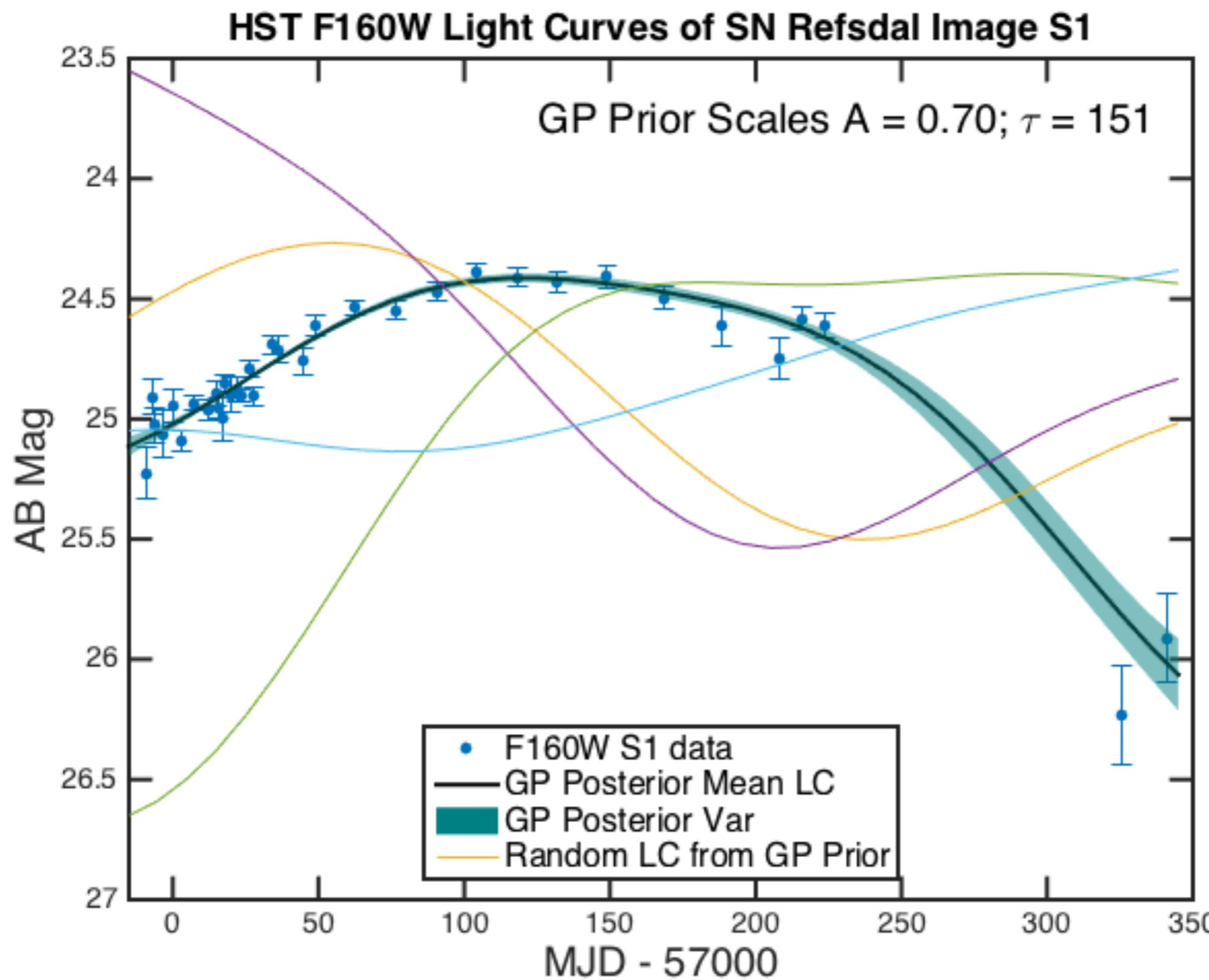
Using Gaussian Conditional Properties:

$$\mathbb{E}[f_* | \mathbf{y}_o] = 1c + \mathbf{K}(t_*, t_o)[\mathbf{K}(t_o, t_o) + \mathbf{W}]^{-1}(\mathbf{y}_o - 1c)$$

$$\text{Var}[f_* | \mathbf{y}_o] = \mathbf{K}(t_*, t_*) - \mathbf{K}(t_*, t_o)[\mathbf{K}(t_o, t_o) + \mathbf{W}]^{-1}\mathbf{K}(t_o, t_*)$$

Posterior Inference with GPs

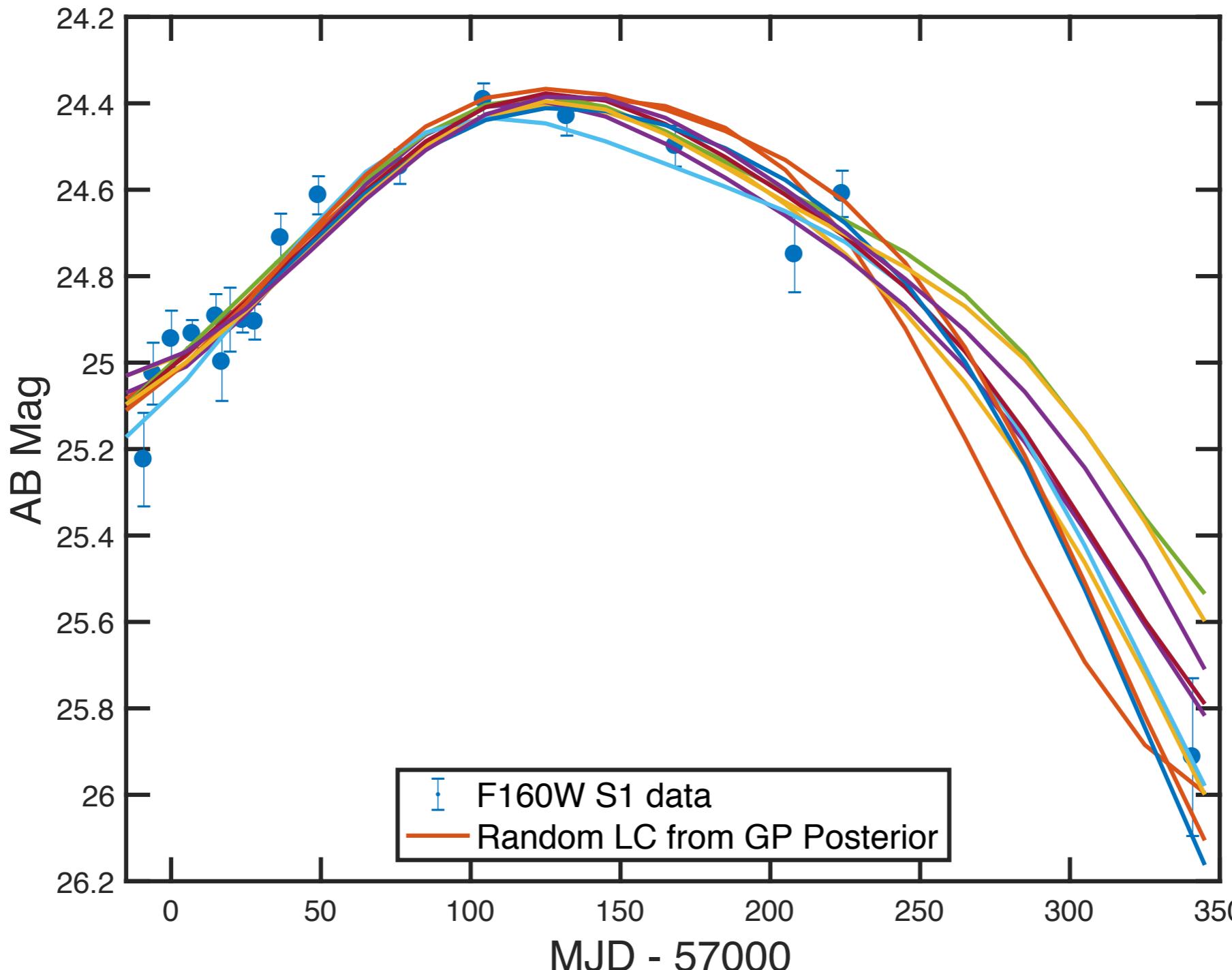
Accounting for Measurement Error



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 , τ^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 , τ)

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

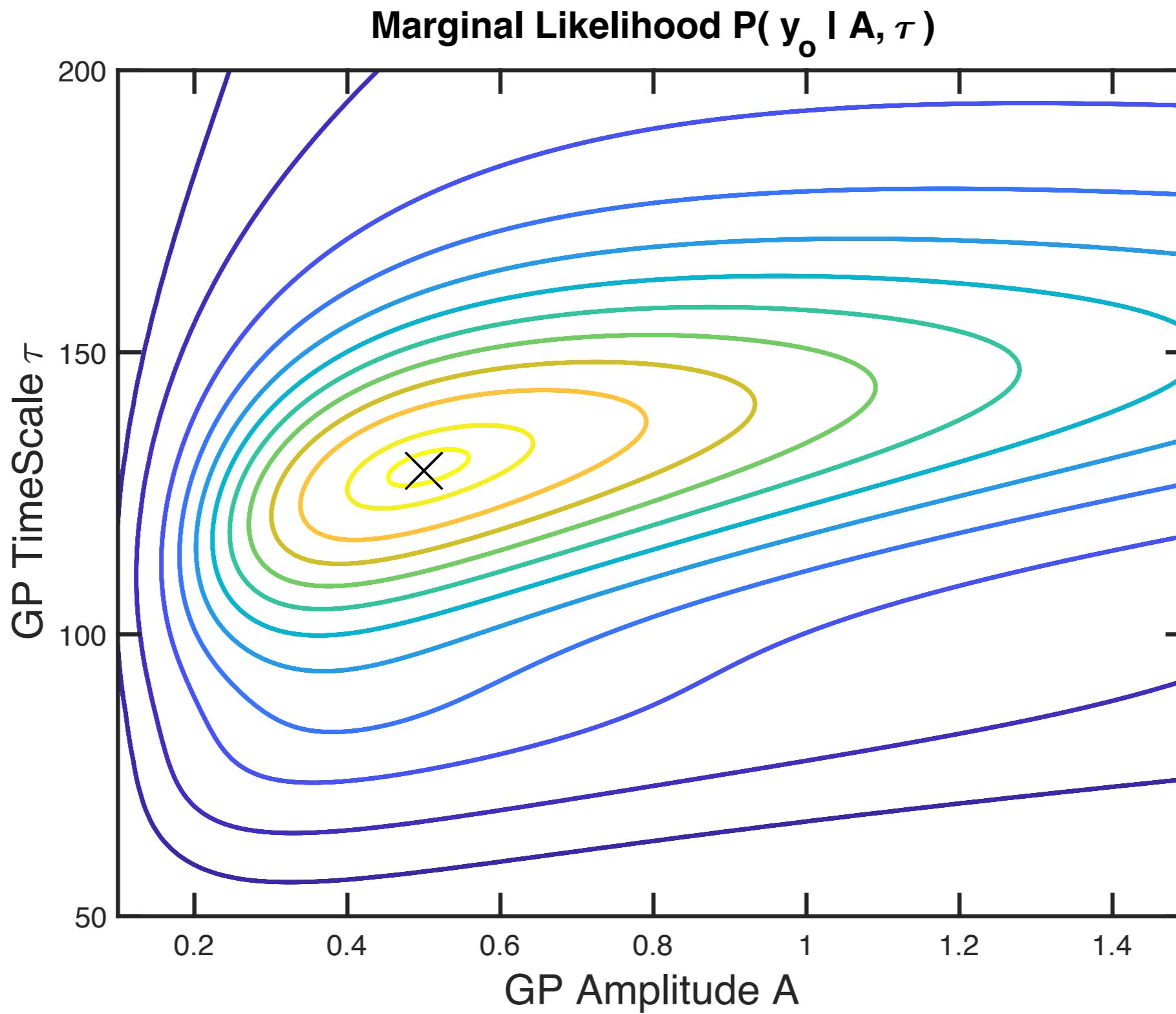
$$L(A, \tau^2) = P(\mathbf{y}_o | A, \tau^2) = N[\mathbf{y}_o | \mathbf{1}c, \mathbf{K}_{A, \tau^2}(\mathbf{t}_o, \mathbf{t}_o) + \mathbf{W}]$$

\mathbf{W} = measurement error covariance

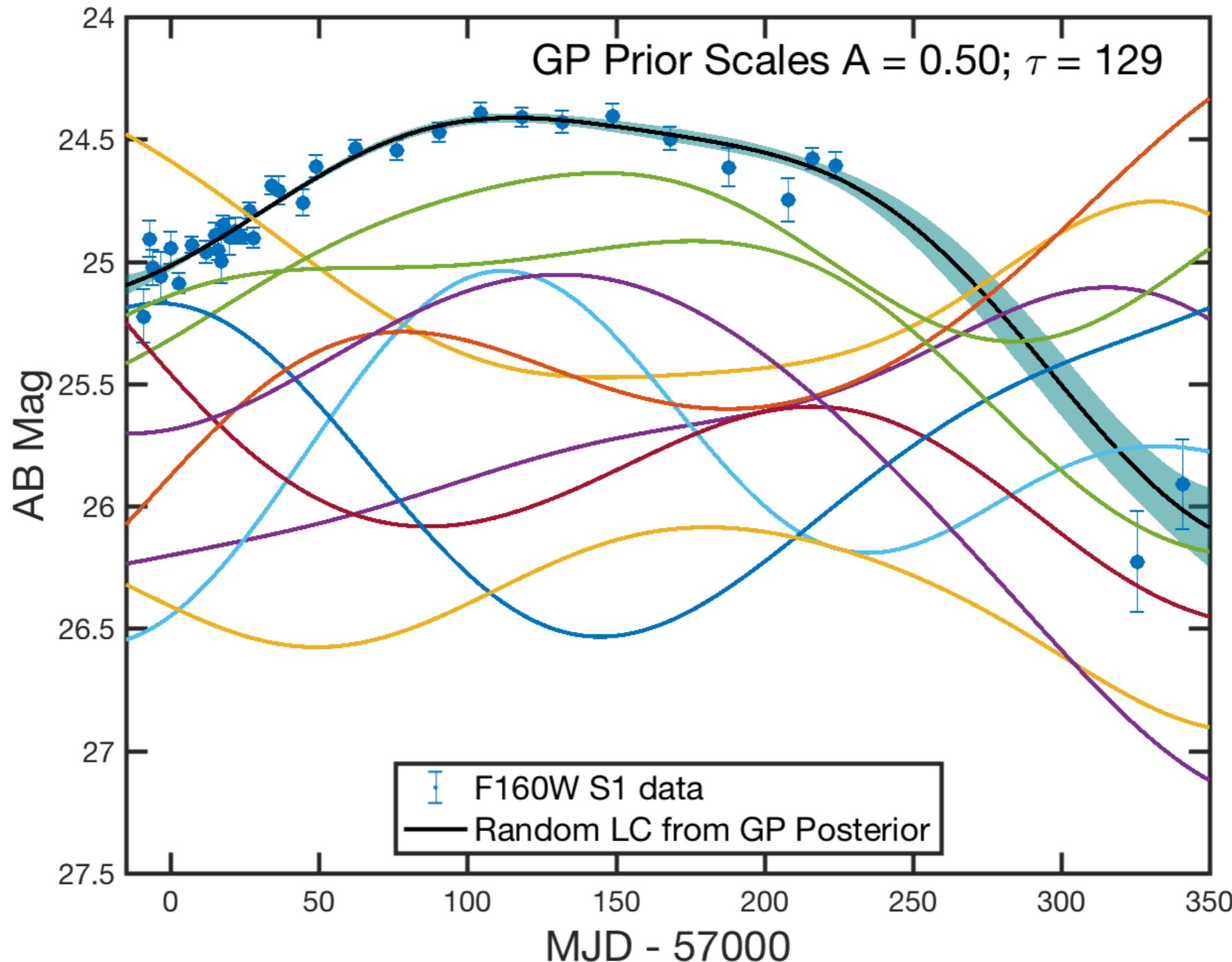
\mathbf{K}_{A, τ^2} = GP covariance

Which we can optimise (max likelihood)
or specify a prior on (A , τ) and sample from posterior

Code Demo



GP Fit with optimised hyperparameters



Other covariance functions (R&W Chapter 4)

Squared Exponential gives very smooth curves.

Ornstein Uhlenbeck Process (Damped Random Walk)

Exponential Covariance Function

$$k(t, t') = A \exp(-|t - t'|/\tau)$$

(Stationary - time translation invariant, Symmetric)

Special Case of Matern kernel:

The Matérn Class of Covariance Functions

The *Matérn class* of covariance functions is given by

$$k_{\text{Matern}}(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}r}{\ell} \right)^\nu K_\nu \left(\frac{\sqrt{2\nu}r}{\ell} \right), \quad (4.14)$$

with positive parameters ν and ℓ , where K_ν is a modified Bessel function

Other covariance functions

Ornstein Uhlenbeck Process (Damped Random Walk)

Exponential Covariance Function

$$k(t, t') = A \exp(-|t - t'|/\tau)$$

