# Astrostatistics: 06 Mar 2020

https://github.com/CambridgeAstroStat/PartIII-Astrostatistics-2020

- Today: Gaussian processes in astrophysics
- Example Class 3:
    - Friday, 13 Mar, 12pm, MR5 (usual lecture time)

# 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

# Review: Properties of Multivariate Gaussians

Full probability density:

$\boldsymbol{\Sigma}$ is positive definite

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

Joint distribution of components:

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

If you observe/know/condition on V:

Conditional dist'n:
$$\boldsymbol{U}|\,\boldsymbol{V} \sim N\left(\mathbb{E}[\boldsymbol{U}|\boldsymbol{V}],\, \mathrm{Var}[\boldsymbol{U}|\boldsymbol{V}]\right)$$

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

Conditional Variance:
$$\mathrm{Var}[\boldsymbol{U}|\boldsymbol{V}] = \boldsymbol{\Sigma}_U - \boldsymbol{\Sigma}_{UV}\,\boldsymbol{\Sigma}_V^{-1}\,\boldsymbol{\Sigma}_{VU}$$

If V = observed data, U = unobserved parameters, then P(**U** | **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.

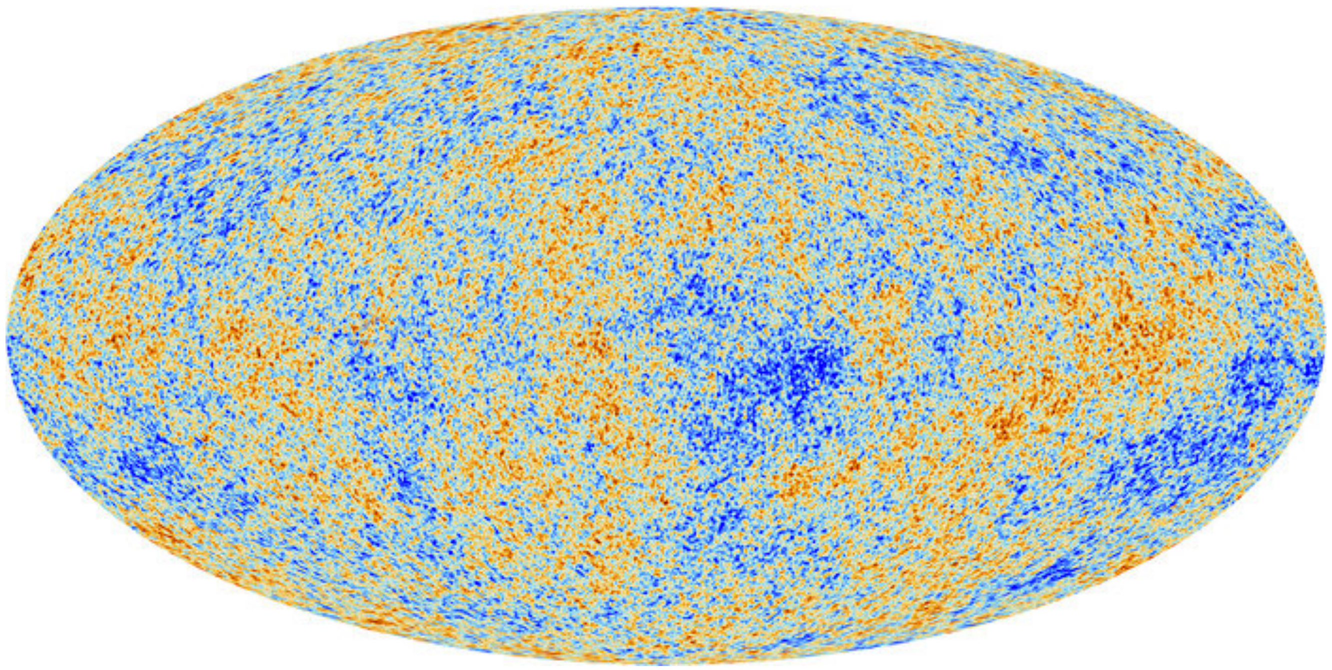
- Any vector $\mathbf{f} = \{f_t : t = 1 \ldots N\}$ of a finite subset is multivariate Gaussian, therefore it is completely described by a mean $\mathbf{E}[\mathbf{f}]$ and covariance matrix $\mathbf{Var}[\mathbf{f}] = \mathbf{Cov}[\mathbf{f}, \mathbf{f}^T]$.

- Elements of the **covariance matrix** are determined by a function of the coordinates, e.g. $\mathrm{Cov}[f_t, f_{t'}] = k(t, t')$, called the *covariance function* or *kernel*

- A Gaussian Process with mean function m(t) is denoted:

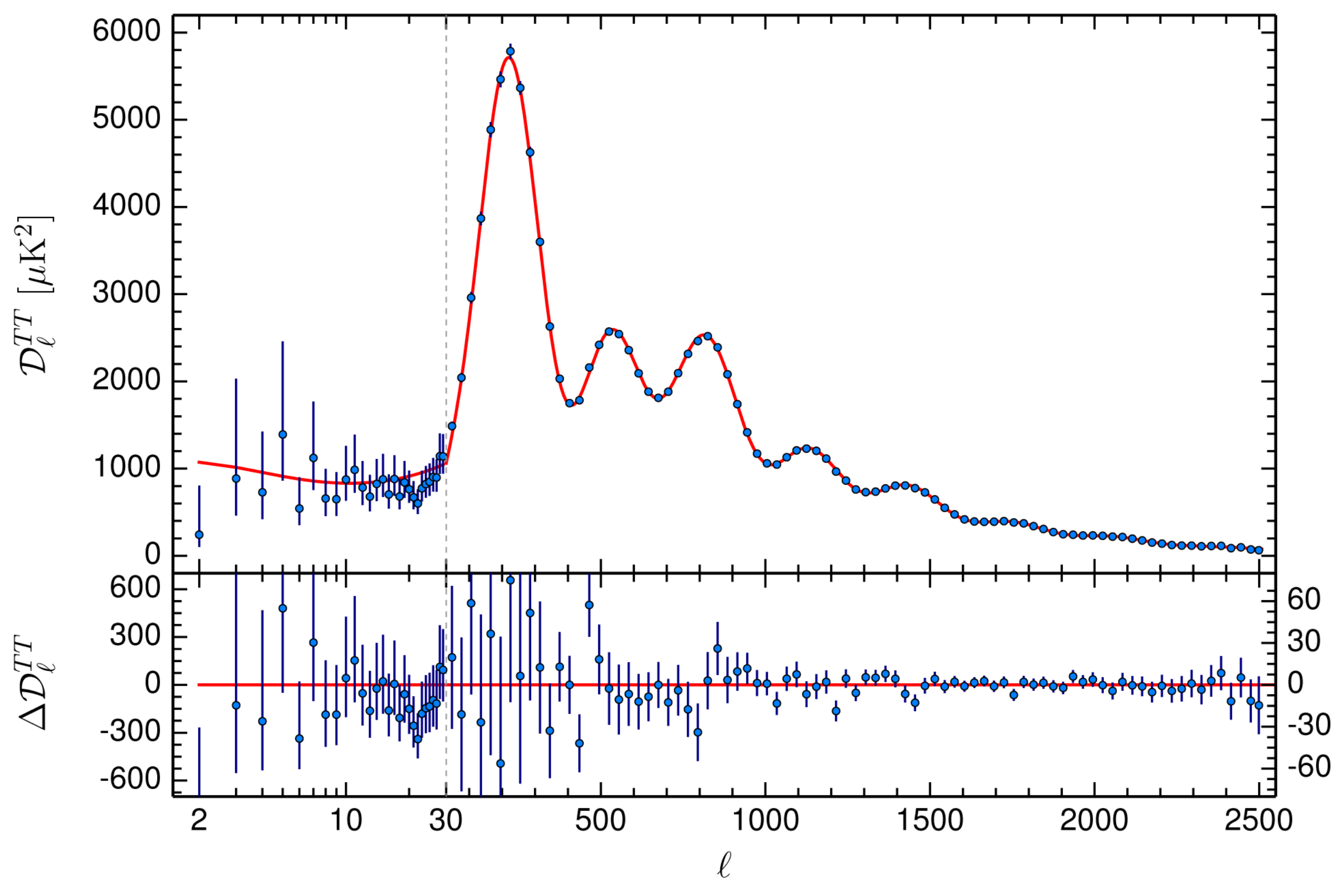
$$f(t) \sim \mathcal{GP}(m(t), k(t, t'))$$

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



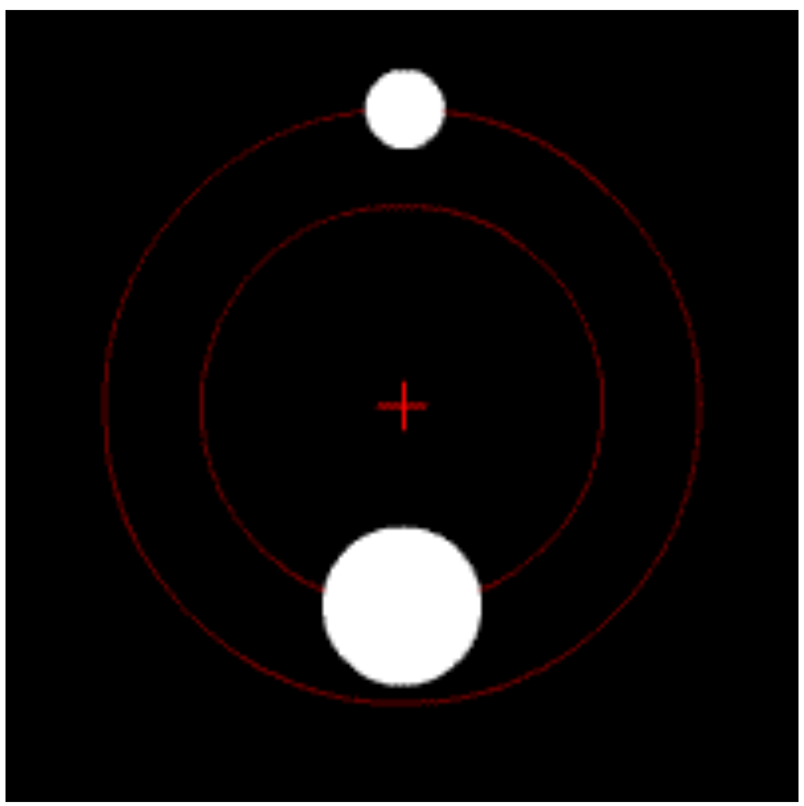
Power Spectrum (~Fourier Transform of Covariance Function)

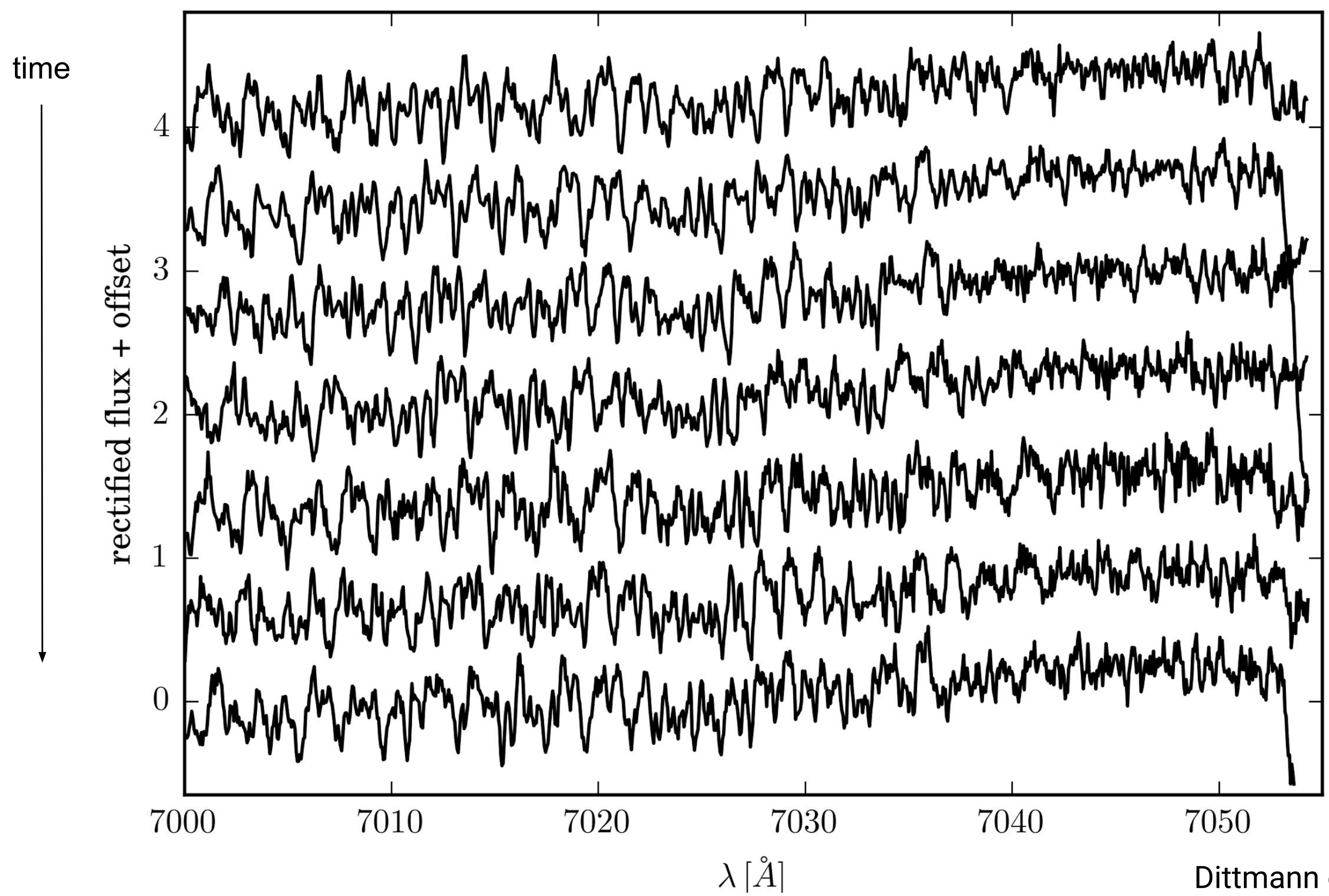
Cosmic Microwave Background (Planck) ~ Gaussian Random Field (mean = 2.7 K, std dev ~ $10^{-5}$ )



Example: *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



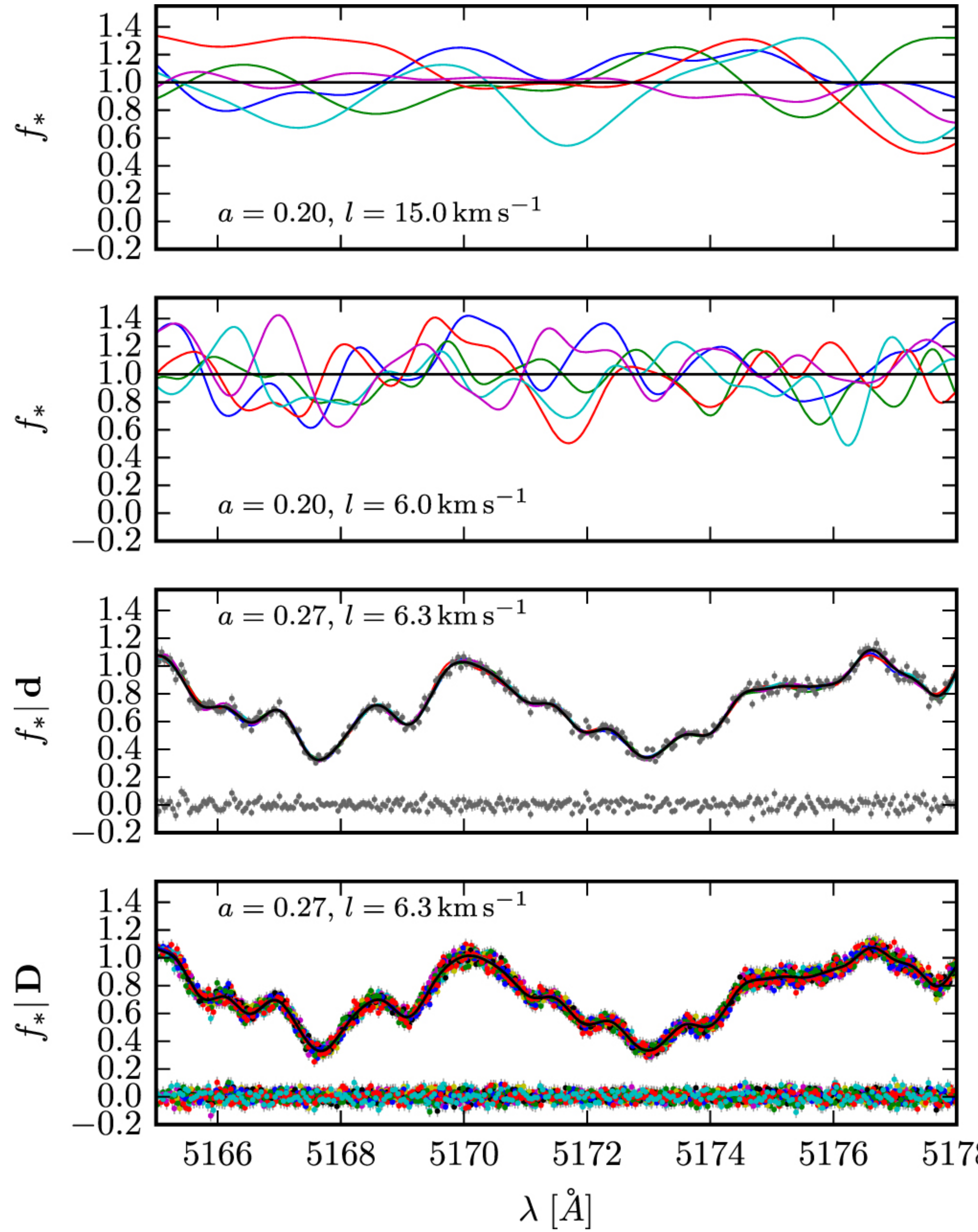



Dittmann et al. 17
**Czekala** et al. 17a

# Example: Gaussian Process: Priors & Posteriors

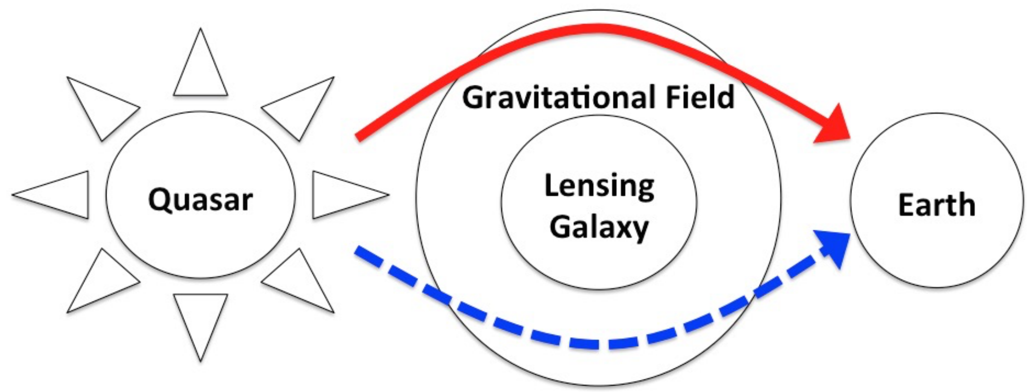
GP prior (long/short correlation length scales)

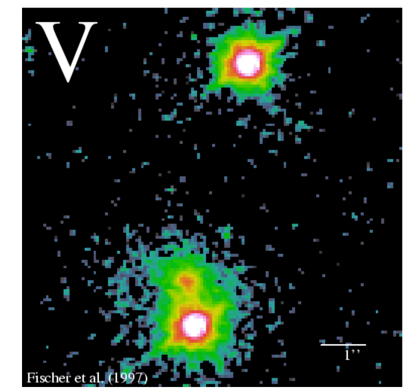
GP Posterior (conditioned on data spectrum d) Inference of latent spectrum



# *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)



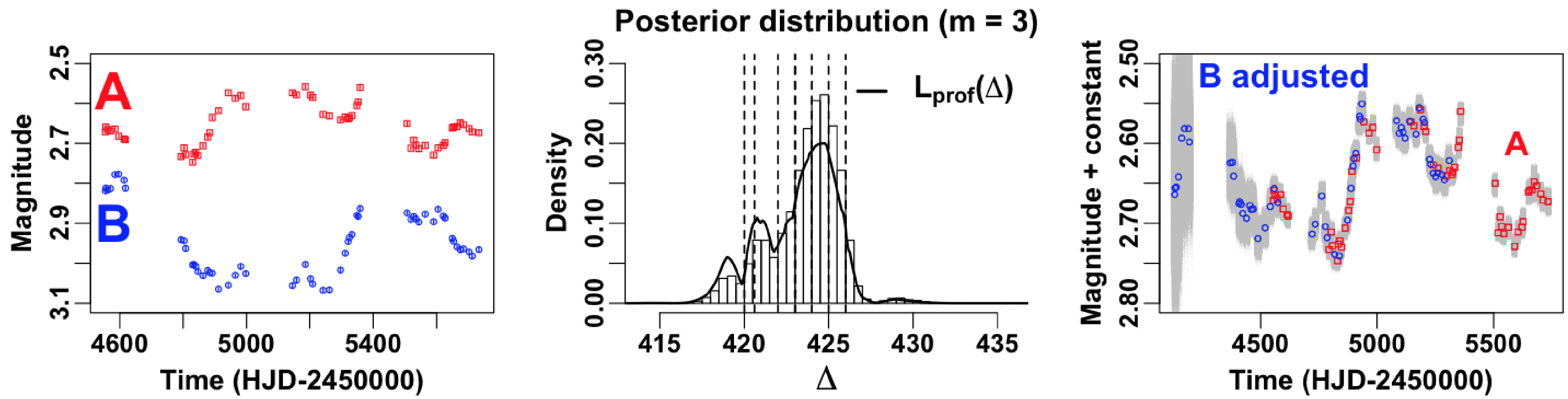
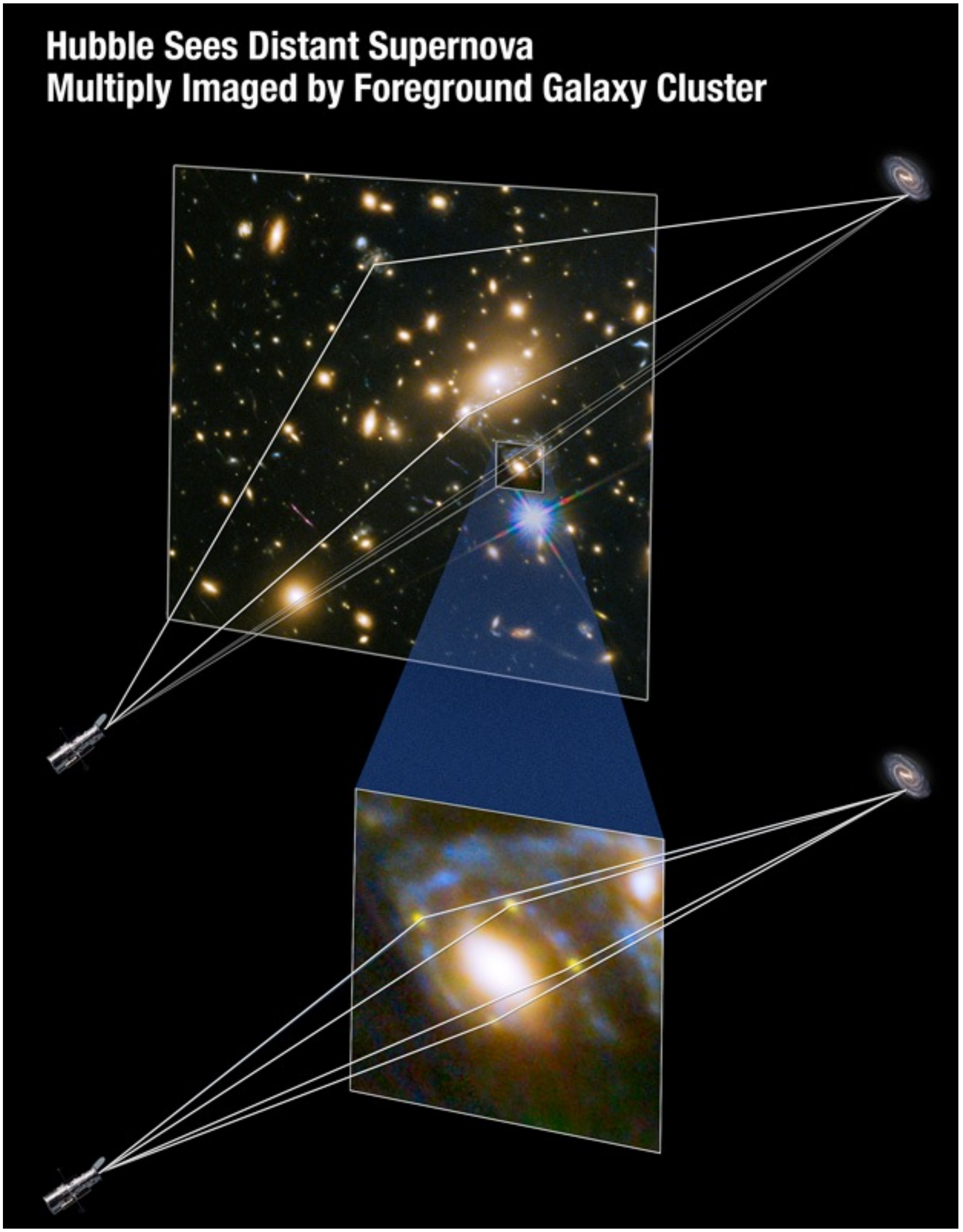
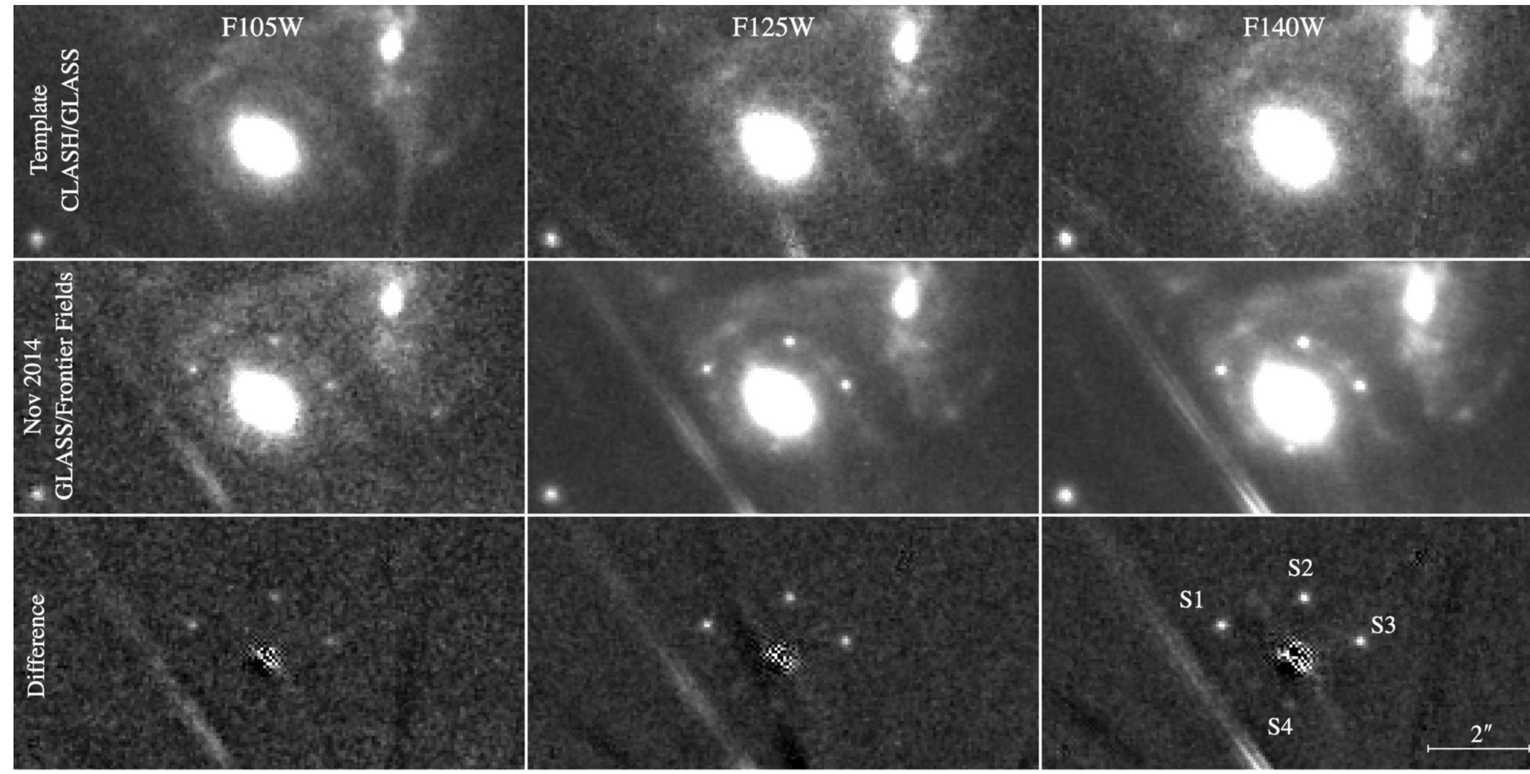



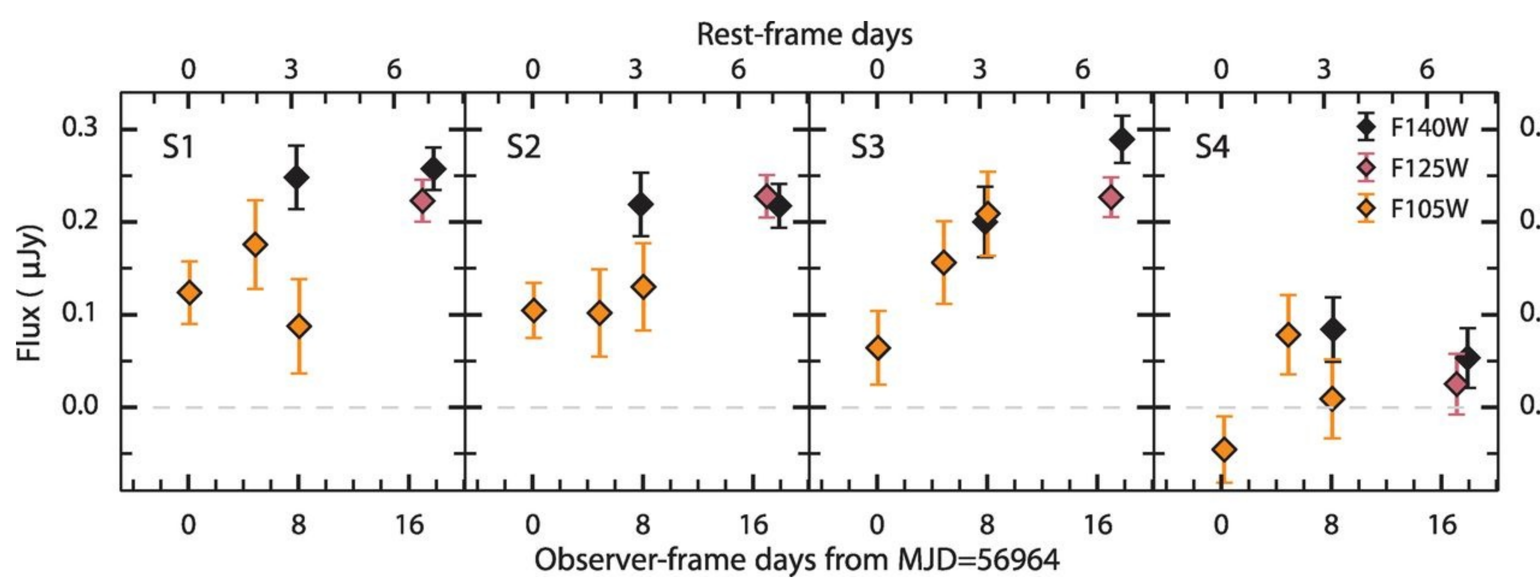

FIG 13. *Observations of Quasar Q0957+561 from Hainline et al. (2012) are plotted in the first panel. The second panel exhibits the marginal posterior distribution of $\Delta$ with*

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

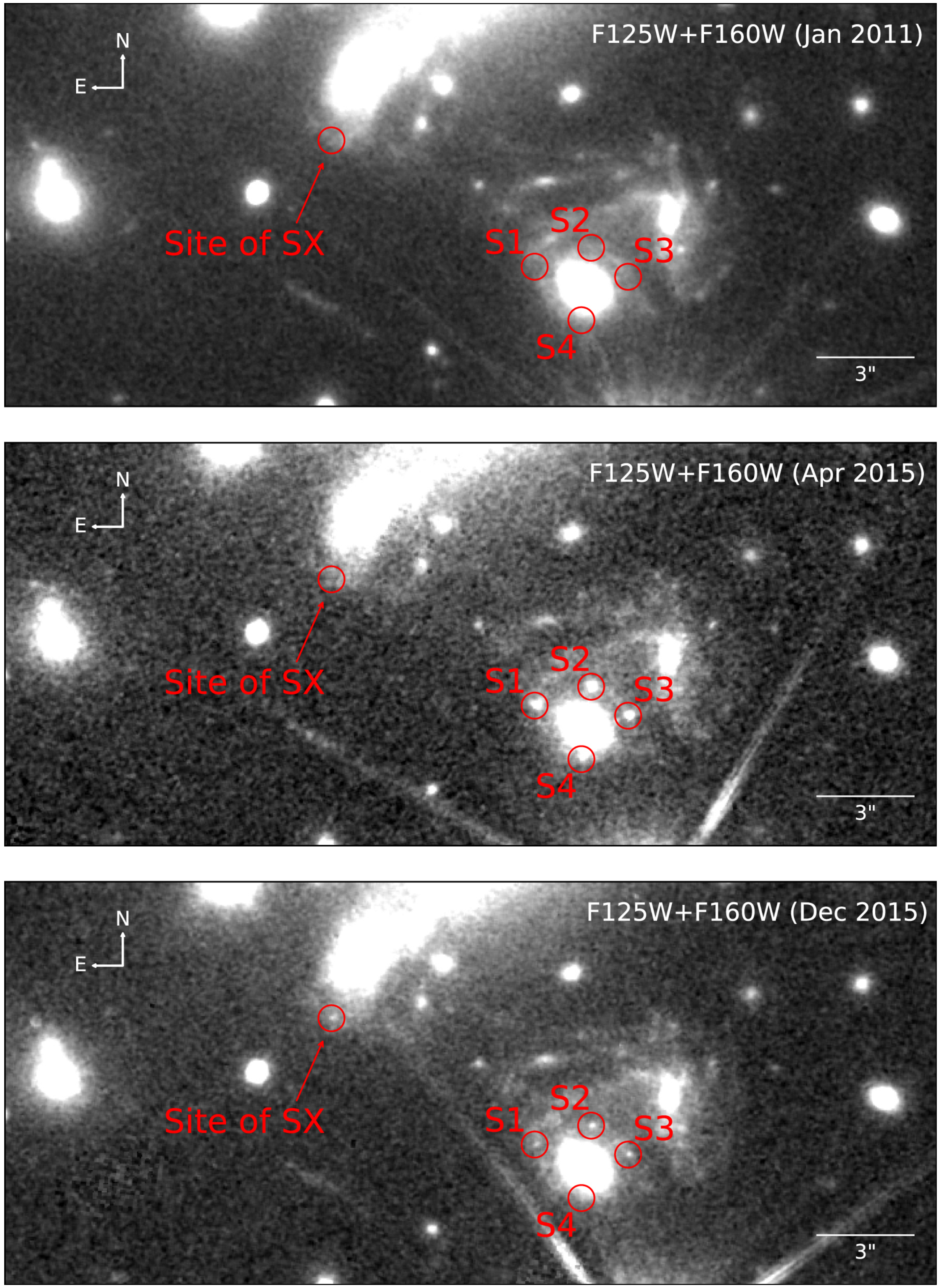


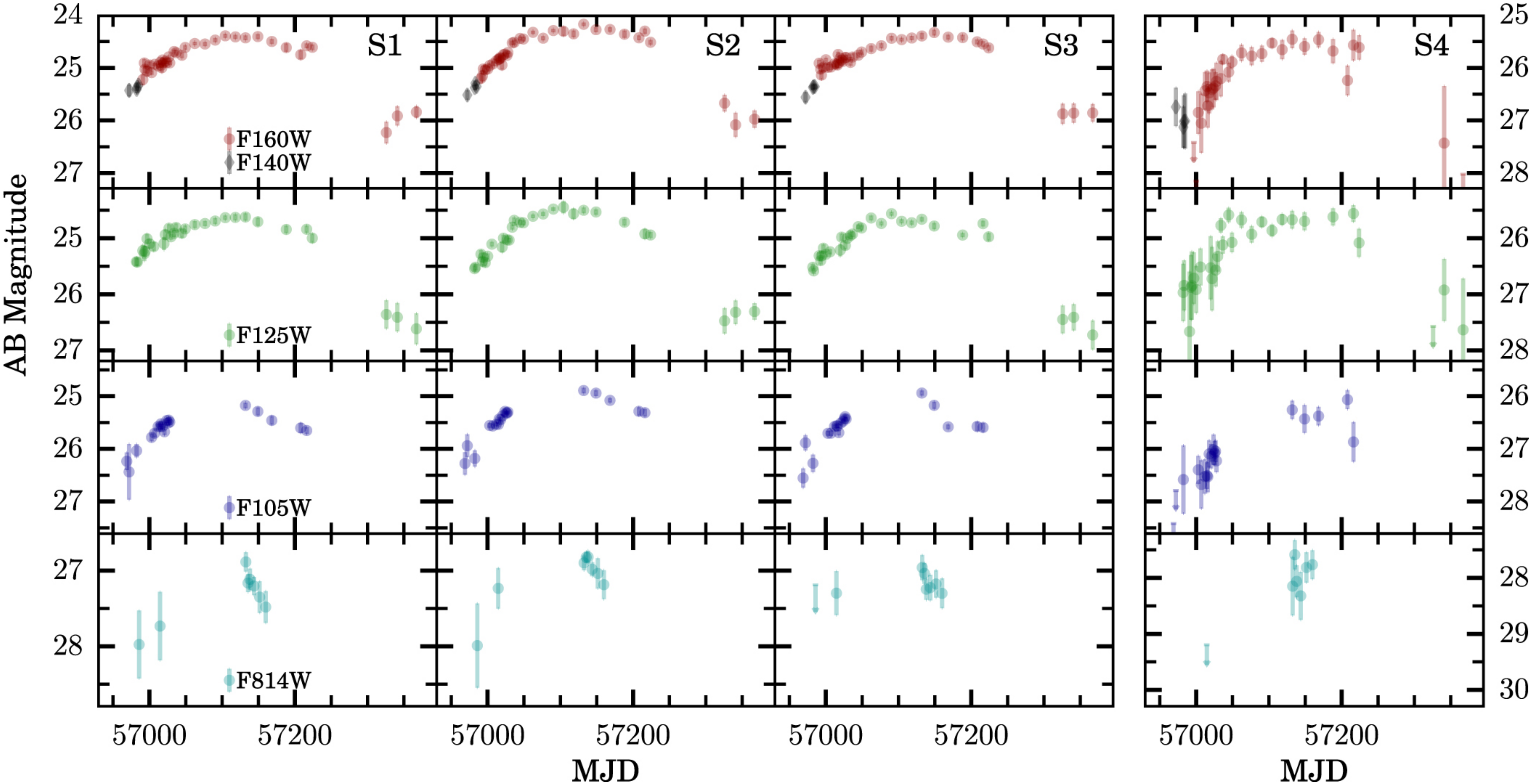




Time Series of SN brightnesses of each image: S1-S4

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

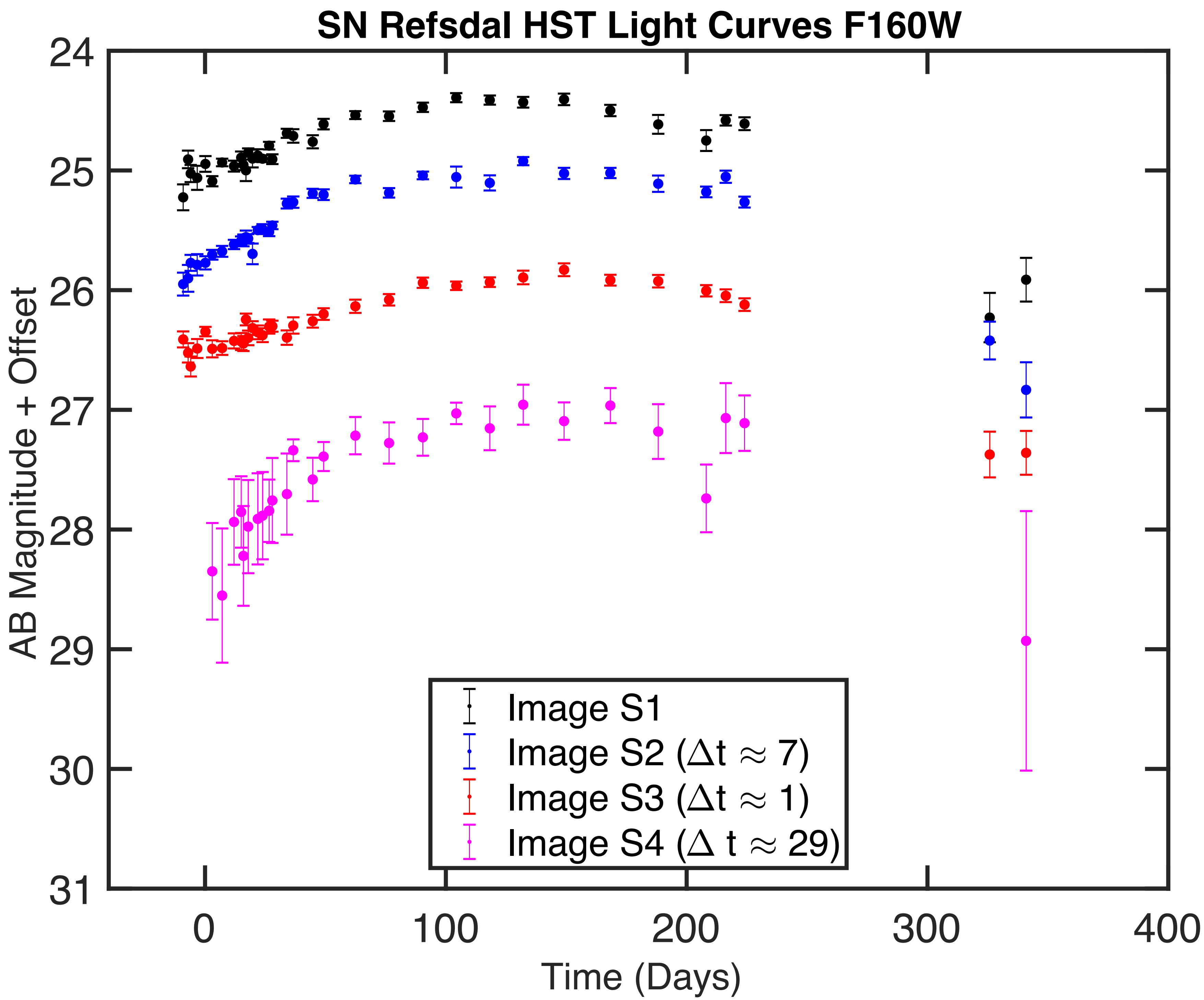


# Hubble Space Telescope time series of SN Refsdal multiple images (Rodney et al. 2016)



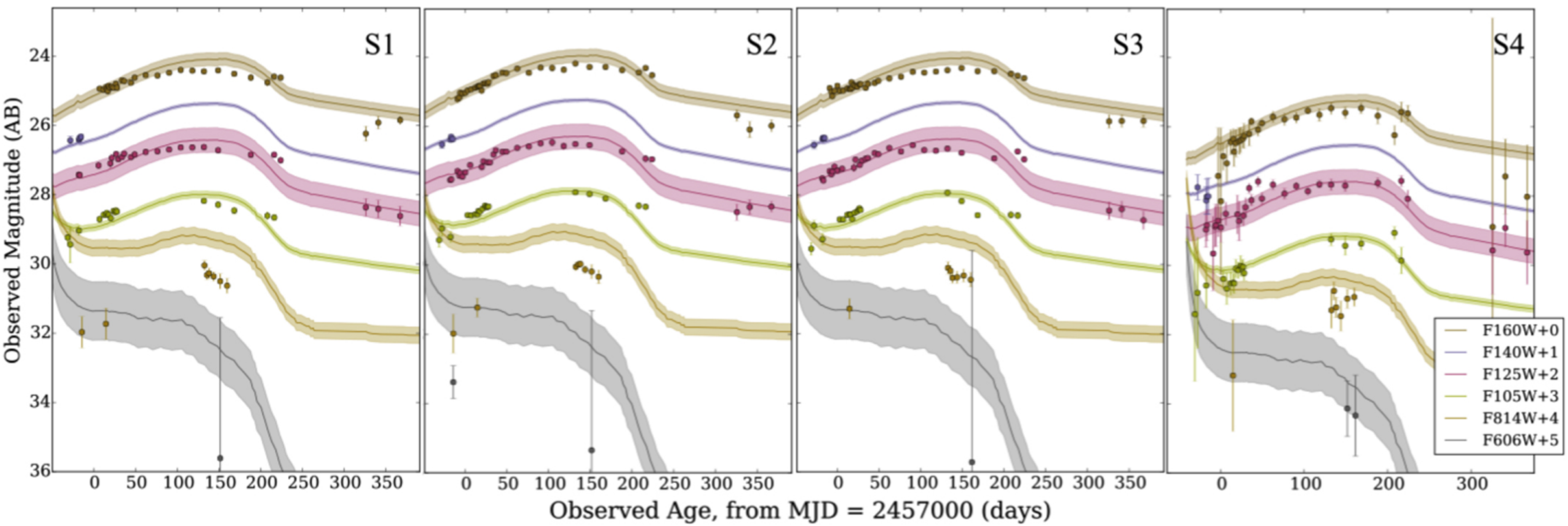

Brightness Time Series [MJD = Modified Julian Day]

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




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)

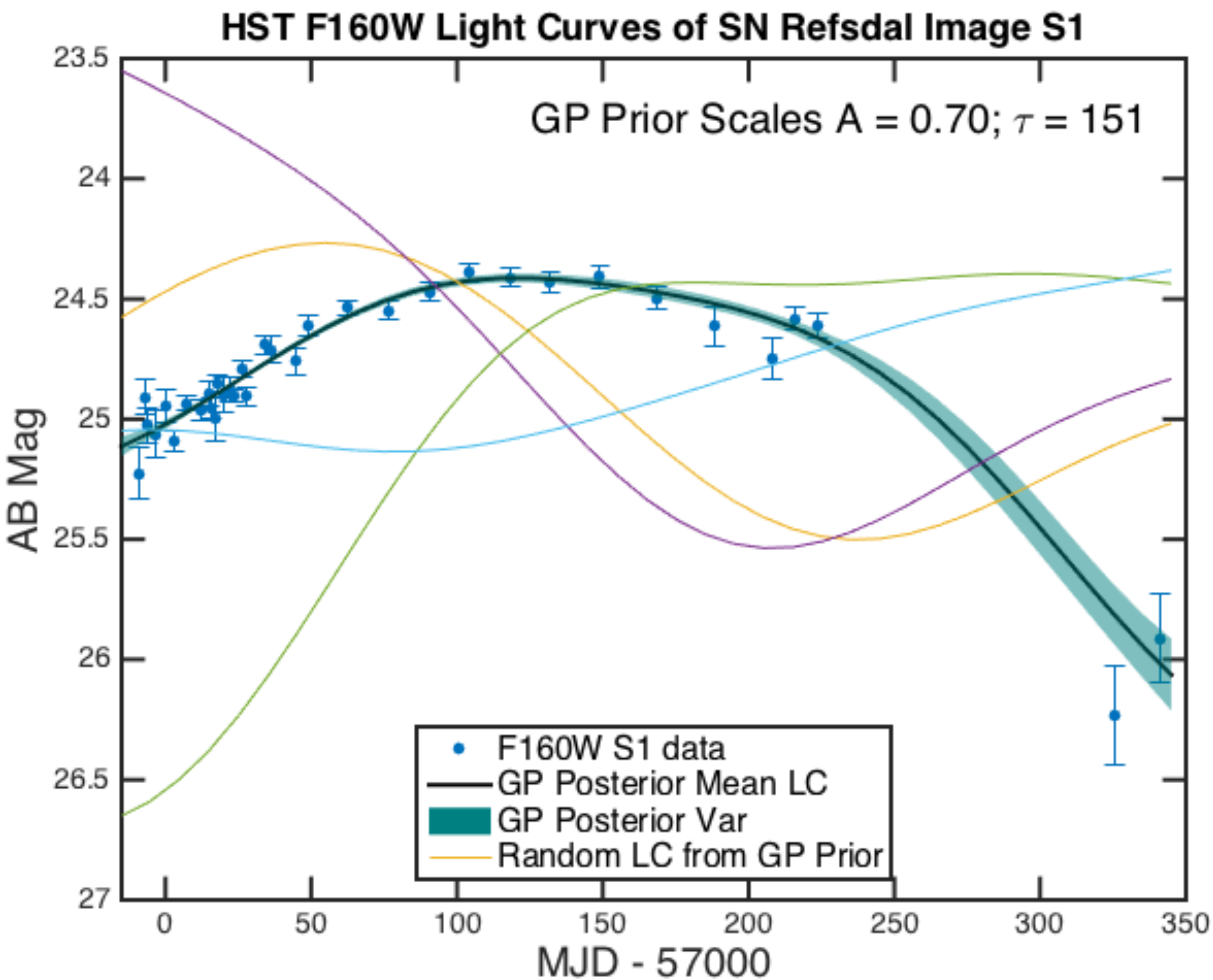



<u>Our Strategy:</u>

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: $\boldsymbol{t} = (t^1, \ldots, t^i, \ldots, t^N)^T$

A vector of function values on the time grid:

$$\boldsymbol{f} = (f(t^1), \ldots, f(t^i), \ldots, f(t^N))^T$$

Assume a Squared Exponential kernel / covariance function

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

Assume a constant prior mean function:

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

# Gaussian Process as a prior on functions

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

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

Covariance Matrix **K** populated by evaluating the kernel:

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

For all pairs of points in **t**:

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

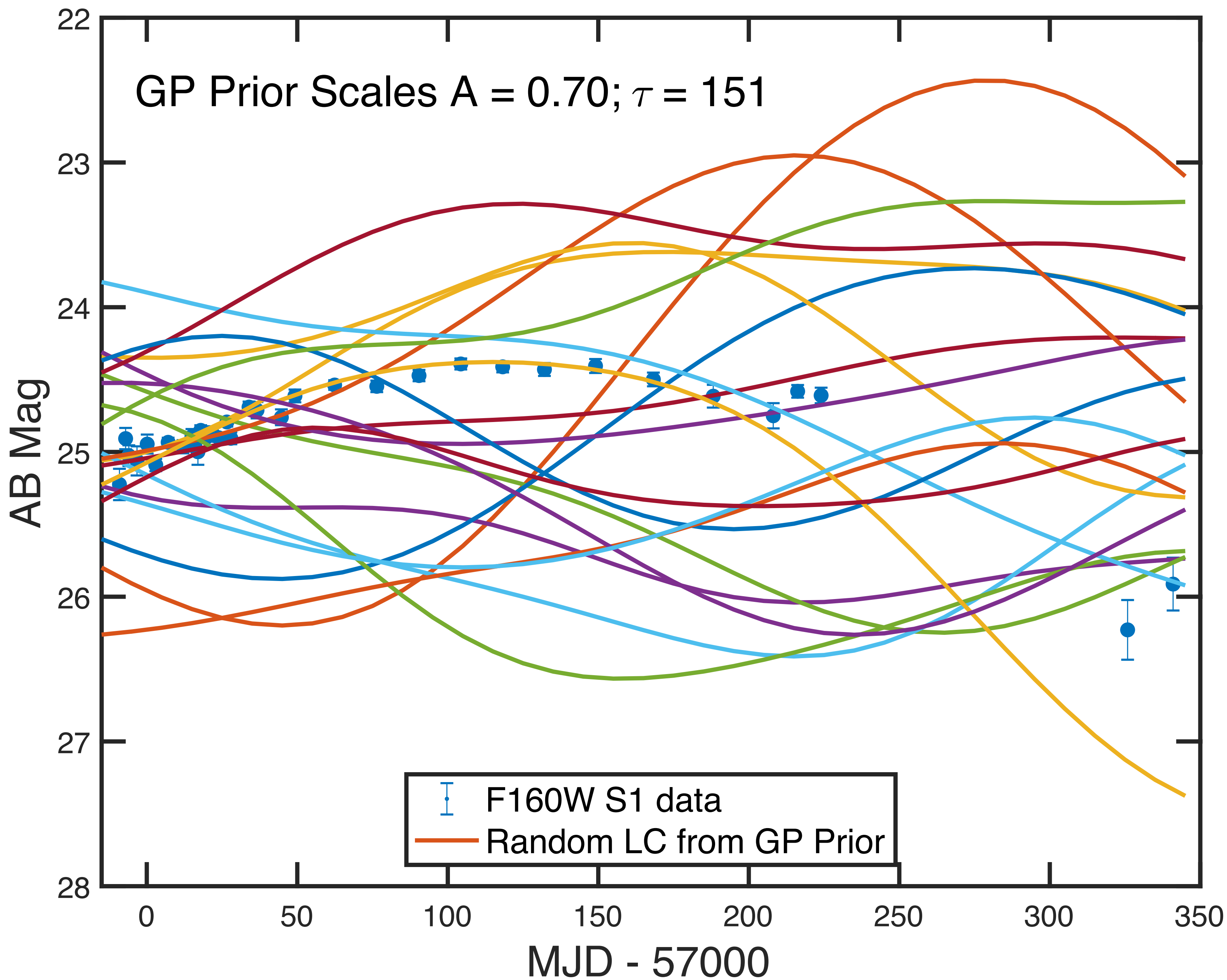
# Drawing random functions from GP prior

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

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

Long Characteristic Timescale

$\tau = 151$



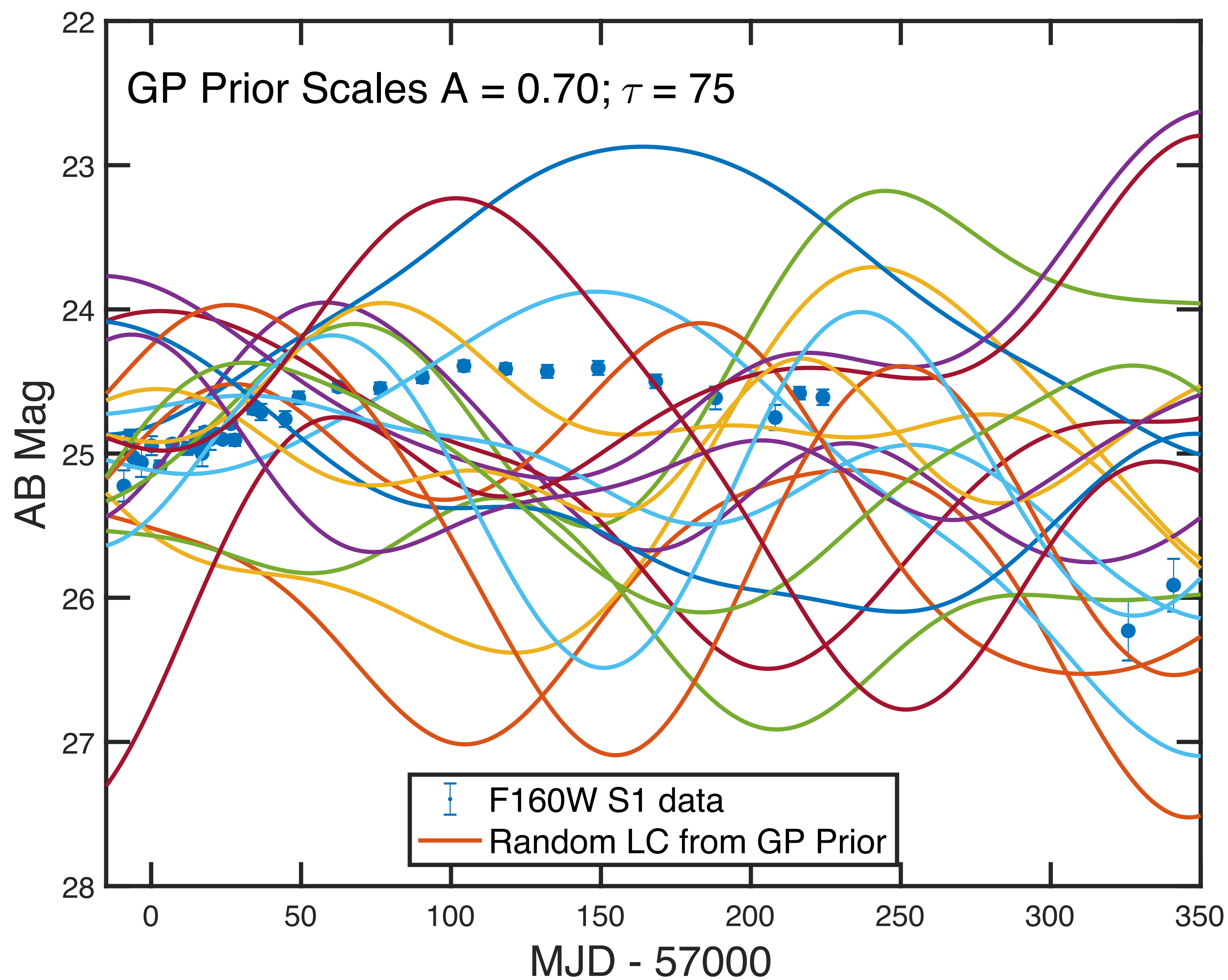
# Drawing random functions from GP prior

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

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

Shorter Characteristic Timescale

$$\tau = 75$$



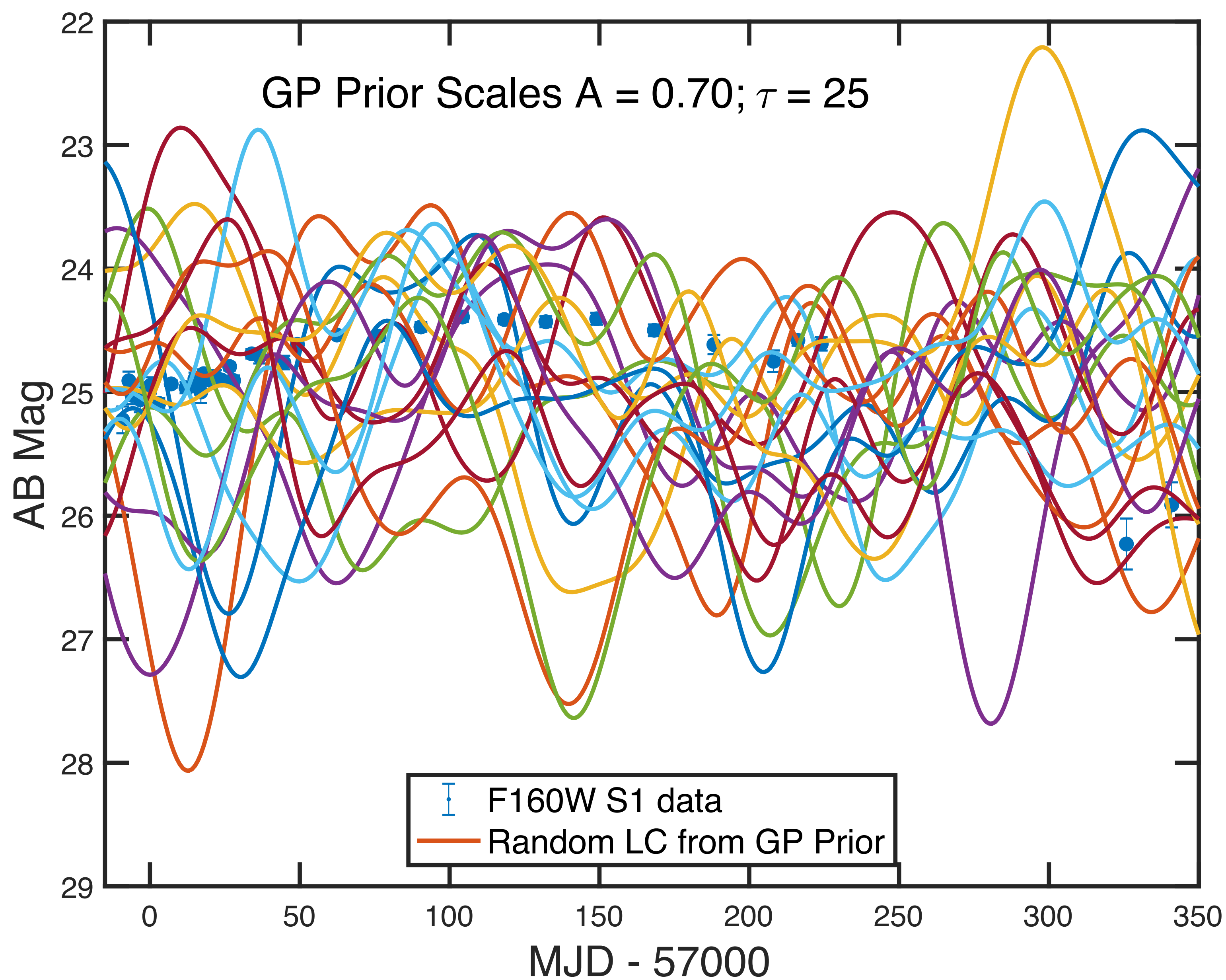
# Drawing random functions from GP prior

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

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

Even Shorter Characteristic Timescale

$$\tau = 25$$



# Fitting a GP to data

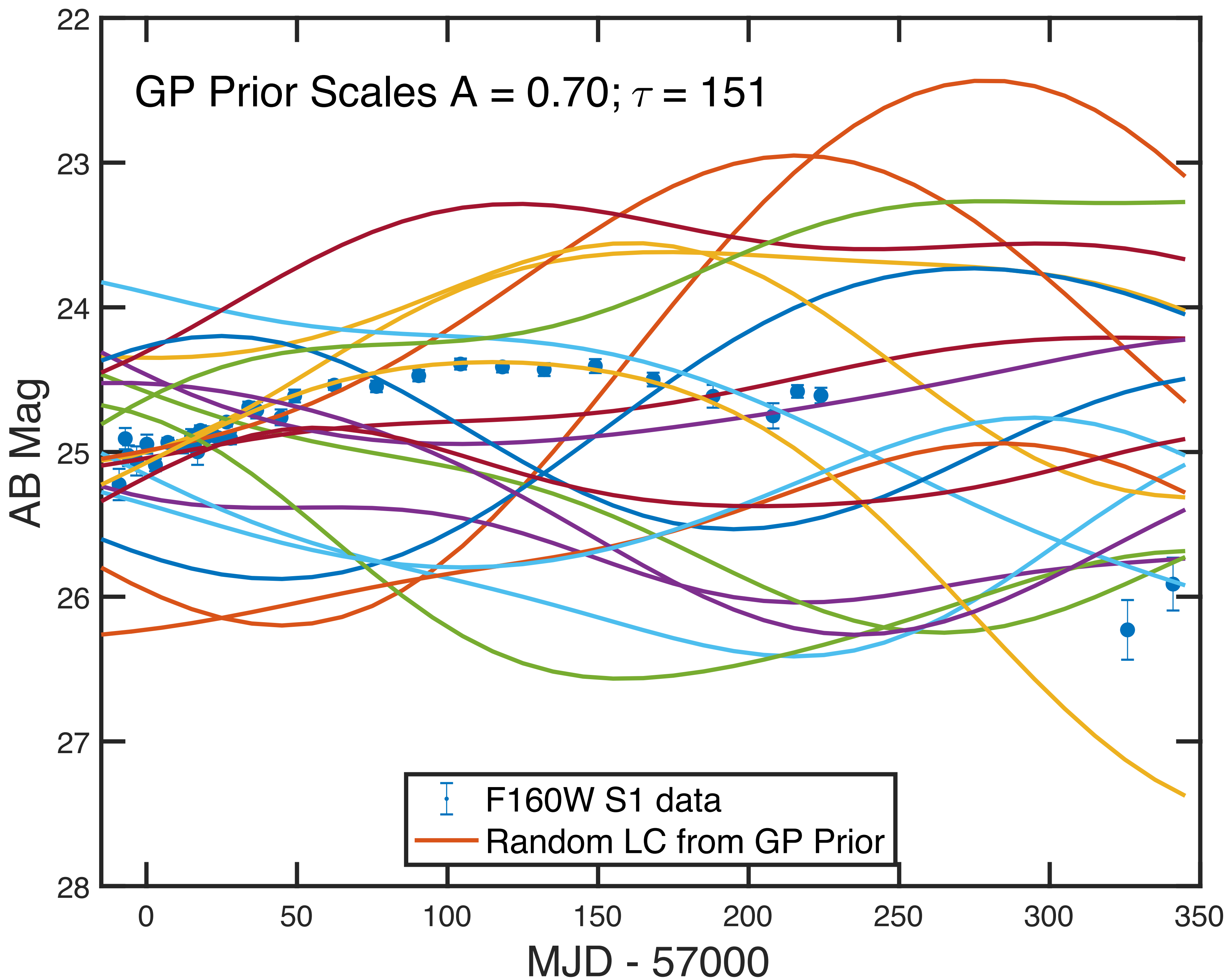
1. If we knew the characteristic scales of the kernel (A, $\boldsymbol{\tau}^2$), then how do we fit the data at observed times to find the curve for unobserved times? (computing the posterior)

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

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

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

Posterior Mean:

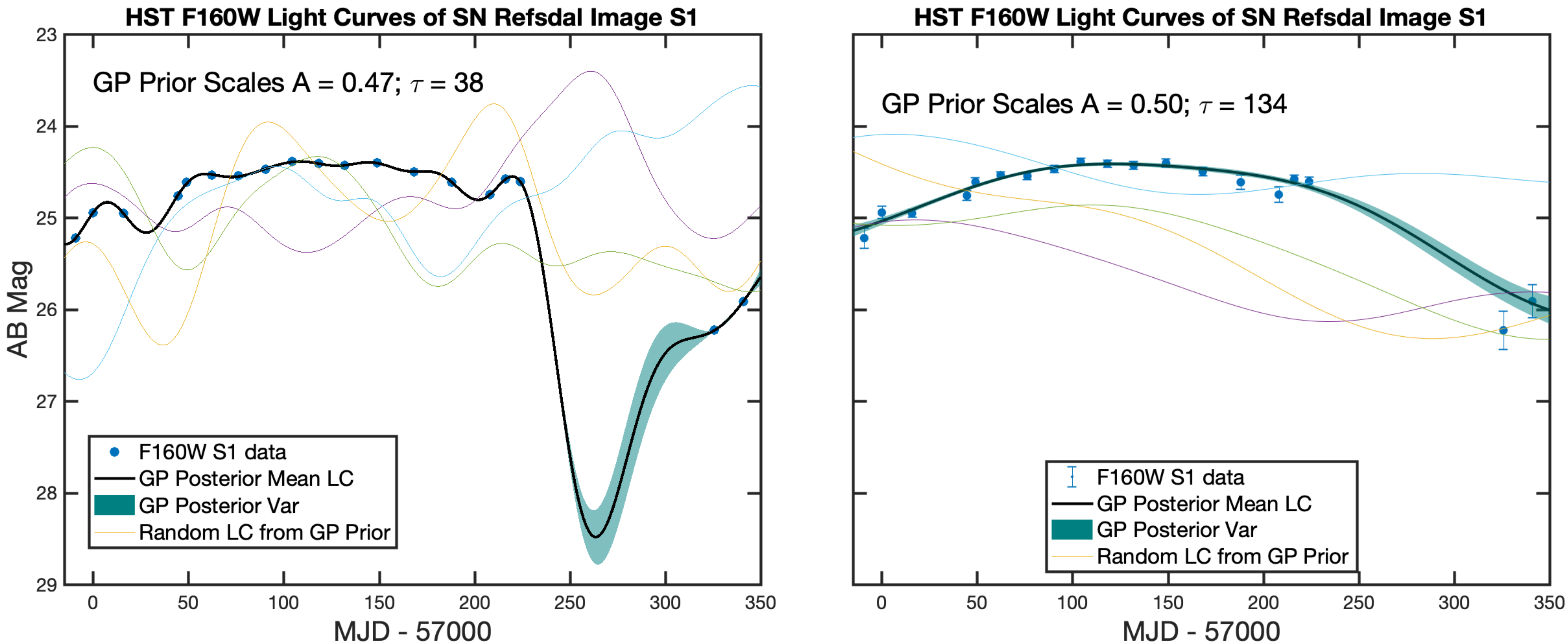
$$\mathbb{E}[\boldsymbol{f}_* | \boldsymbol{f}_o] = \mathbf{1}c + \boldsymbol{K}(\boldsymbol{t}_*, \boldsymbol{t}_o)\boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_o)^{-1}(\boldsymbol{f}_o - \mathbf{1}c)$$

Posterior Co(variance):

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

# Posterior Inference with GPs
# Estimating the underlying curve:

## Ignore measurement error    Include measurement error




Need to:

1. Account for Measurement Error
2. Adapt Hyperparameters

# 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)

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

<u>Joint Distribution of data and latent function</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: 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:

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

# 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: Another Derivation using Conditional/Marginal properties of MVN

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

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

(Cond & Marg) P(V) & P(U|V) —> P(U) (Marginal)

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

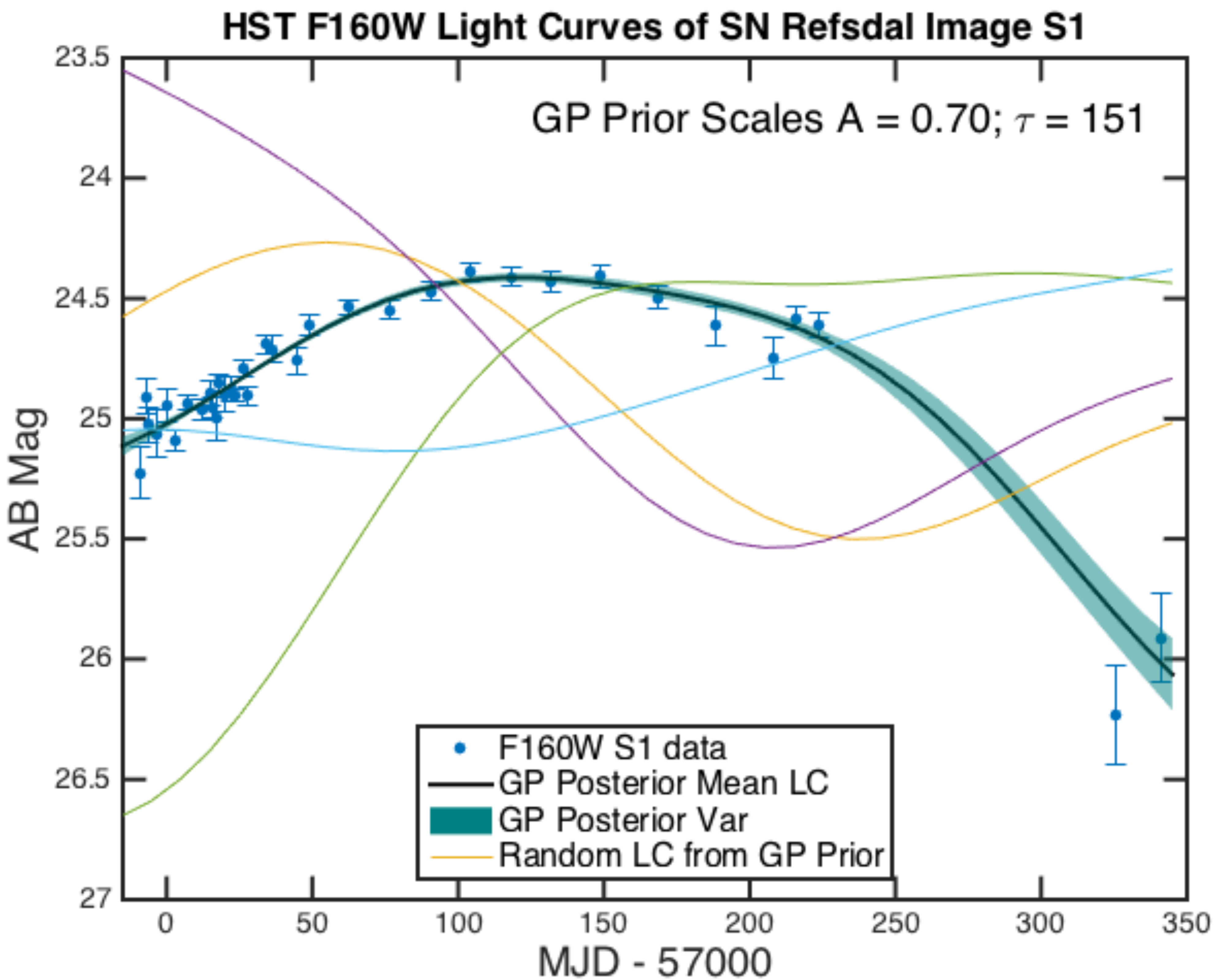
$$\boldsymbol{f}_* | \boldsymbol{y}_o \sim N(\mathbb{E}[\boldsymbol{f}_* | \boldsymbol{y}_o], \mathrm{Var}[\boldsymbol{f}_* | \boldsymbol{y}_o])$$

Using Gaussian Conditional Properties:

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

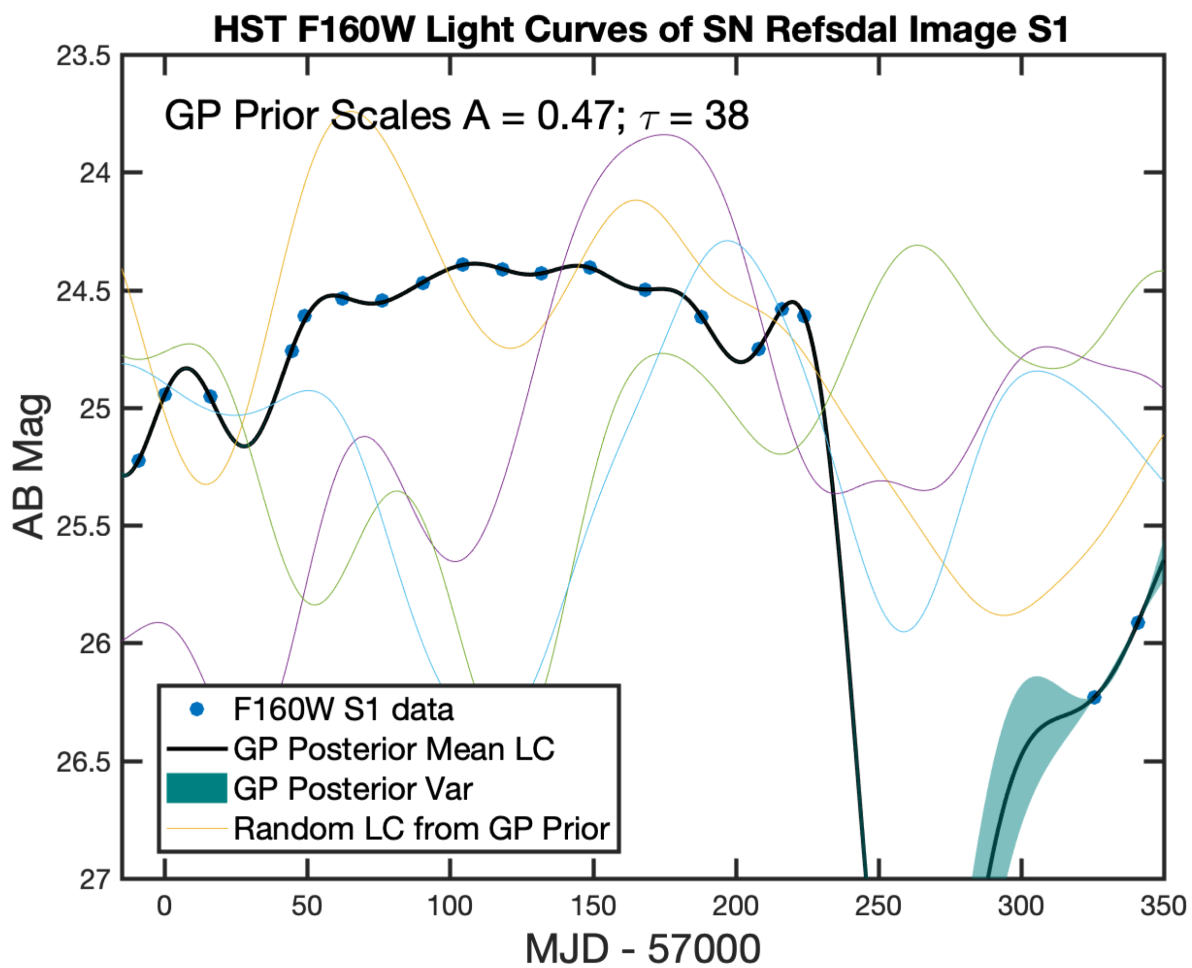
$$\mathrm{Var}[\boldsymbol{f}_* | \boldsymbol{y}_o] = \boldsymbol{K}(\boldsymbol{t}_*, \boldsymbol{t}_*) - \boldsymbol{K}(\boldsymbol{t}_*, \boldsymbol{t}_o)[\boldsymbol{K}(\boldsymbol{t}_{\boldsymbol{o}}, \boldsymbol{t}_o) + \boldsymbol{W}]^{-1}\boldsymbol{K}(\boldsymbol{t}_, \boldsymbol{t}_*)$$

# Posterior Inference with GPs Accounting for Measurement Error



# GP Fitting with measurement error

## (over)fitting $f_0$

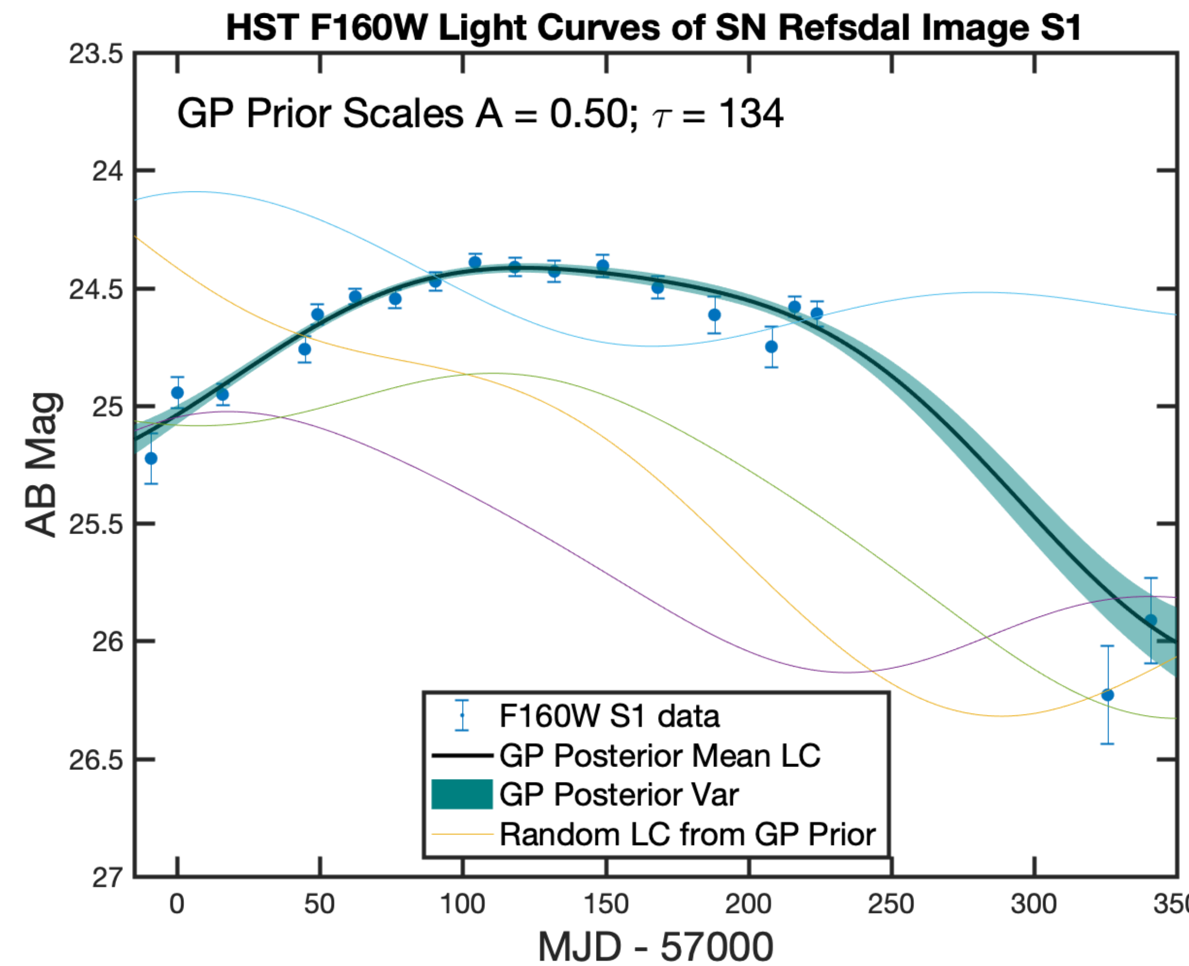



Actually a small “nugget term” was added to regularise covariance matrix inversion.

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

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

## Fitting $y_0$

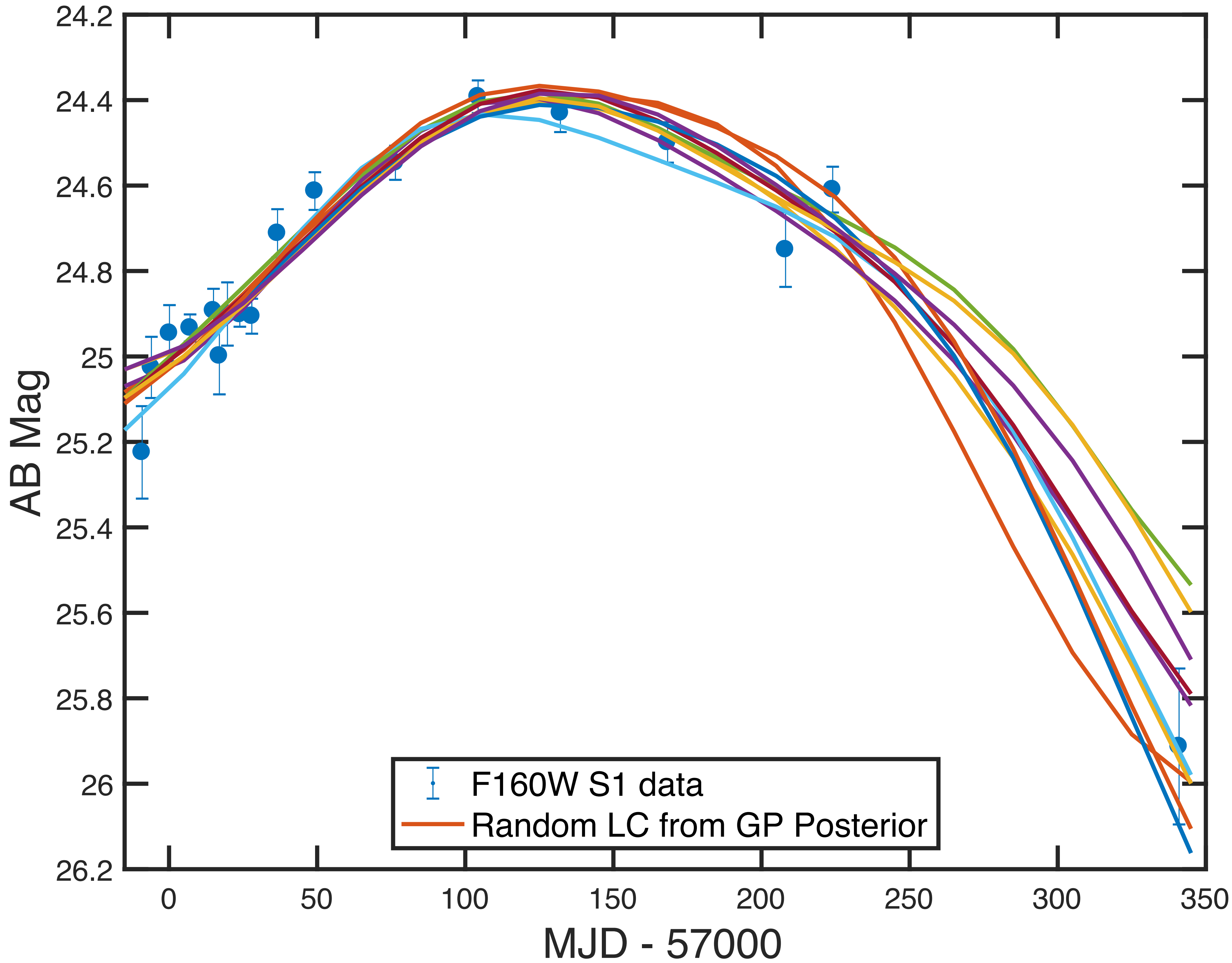



Using actual measurement uncertainties

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

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

## 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:

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

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

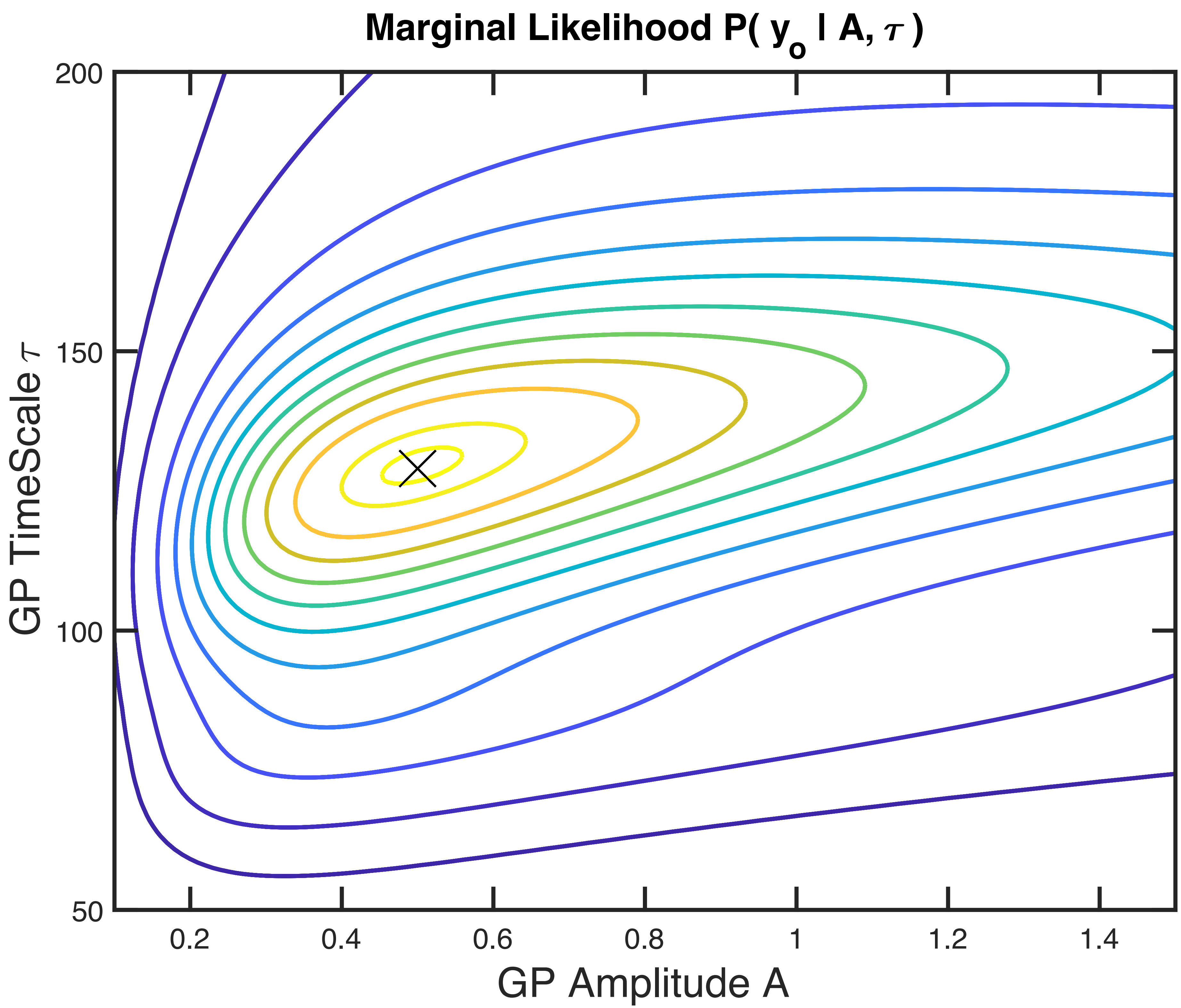
$$L(A, \tau^2) = P(\boldsymbol{y}_o|A, \tau^2) = N[\boldsymbol{y_o}|\, \mathbf{1}c, \boldsymbol{K}_{A,\tau^2}(\boldsymbol{t}_o, \boldsymbol{t}_o) + \boldsymbol{W}]$$

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

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

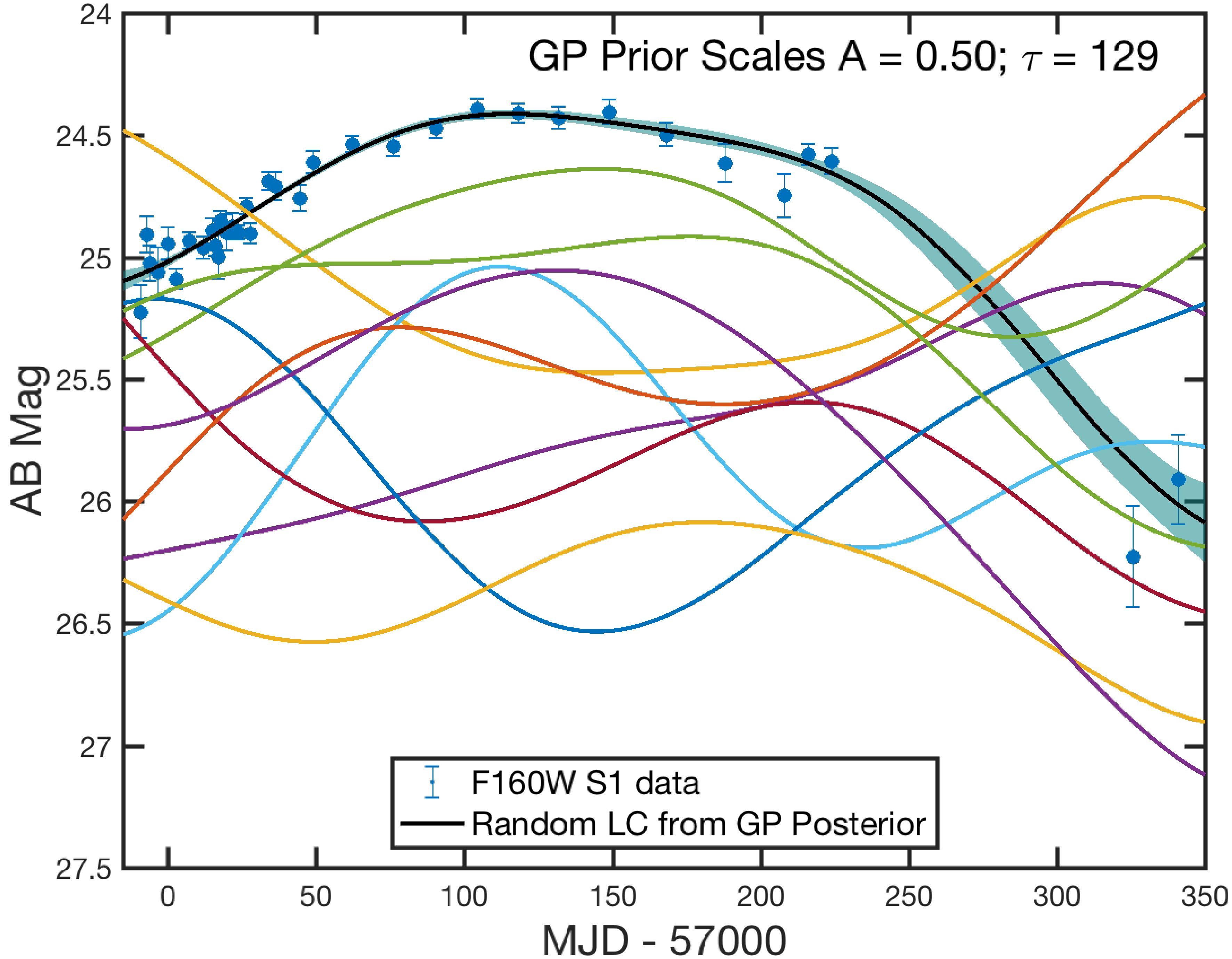
Which we can optimise (max likelihood)
or specify a prior on (A, $\boldsymbol{\tau}$) and sample from posterior

# Inferring Hyperparmeters

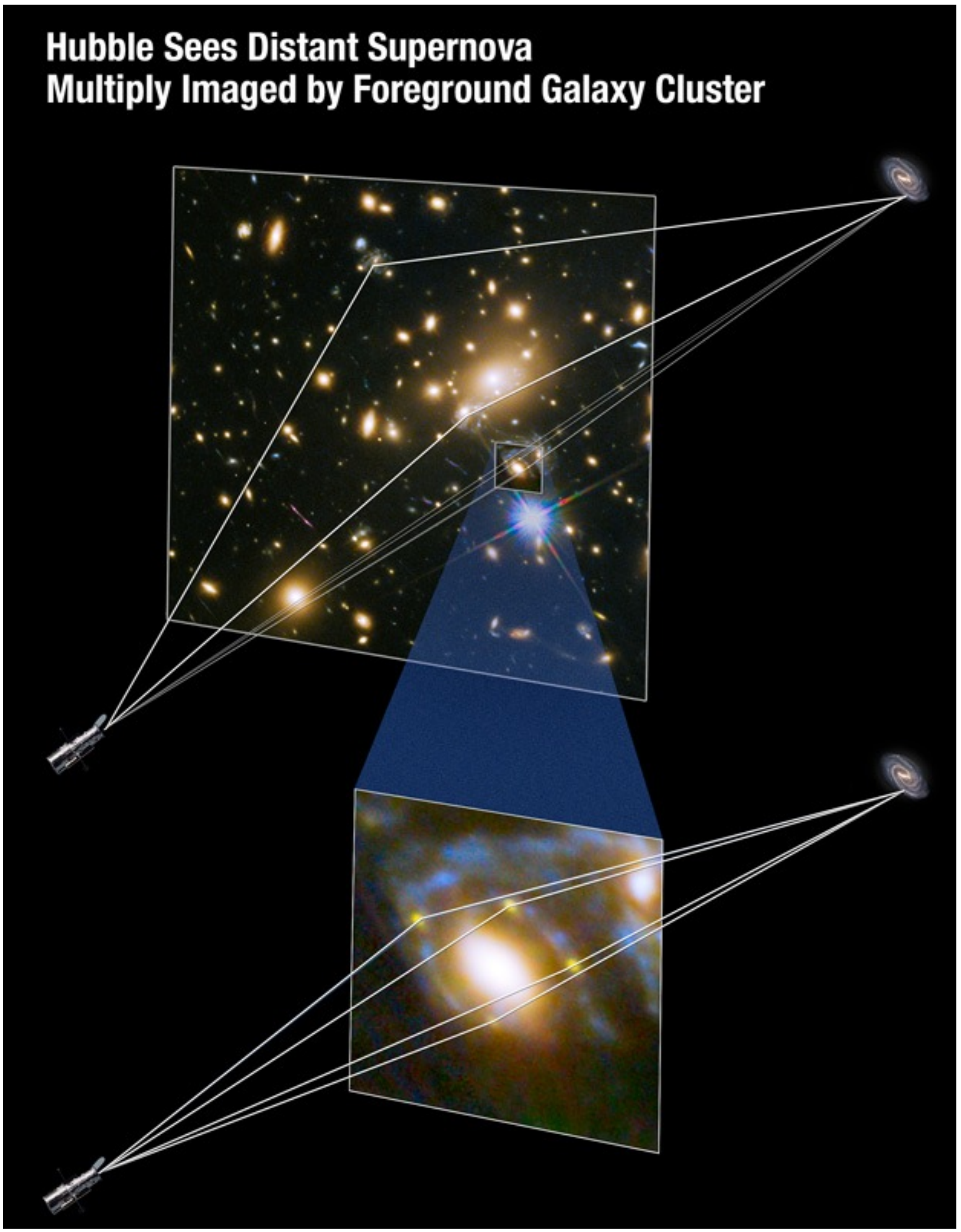


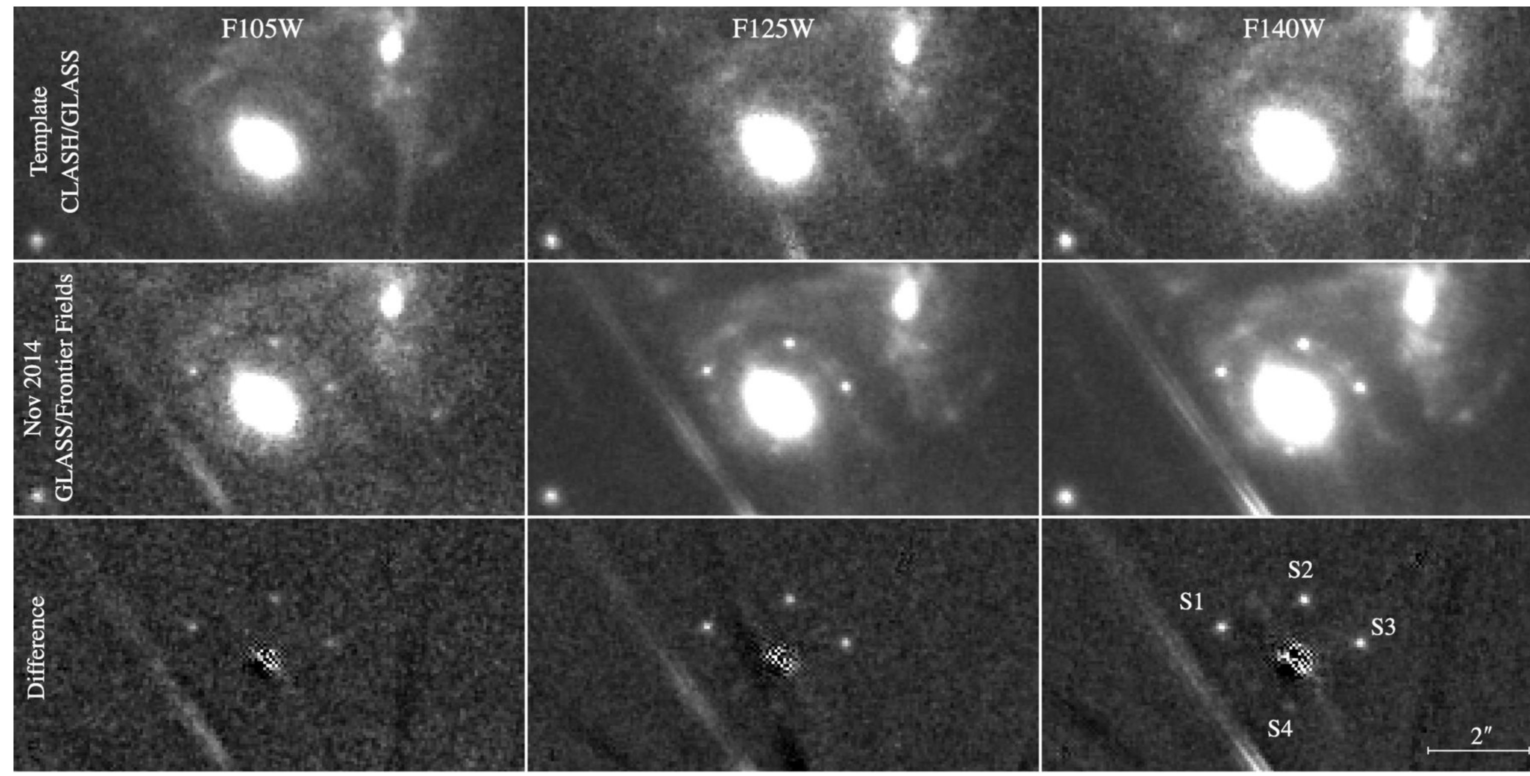

MLE: A = 0.50, $\boldsymbol{\tau}$ = 129

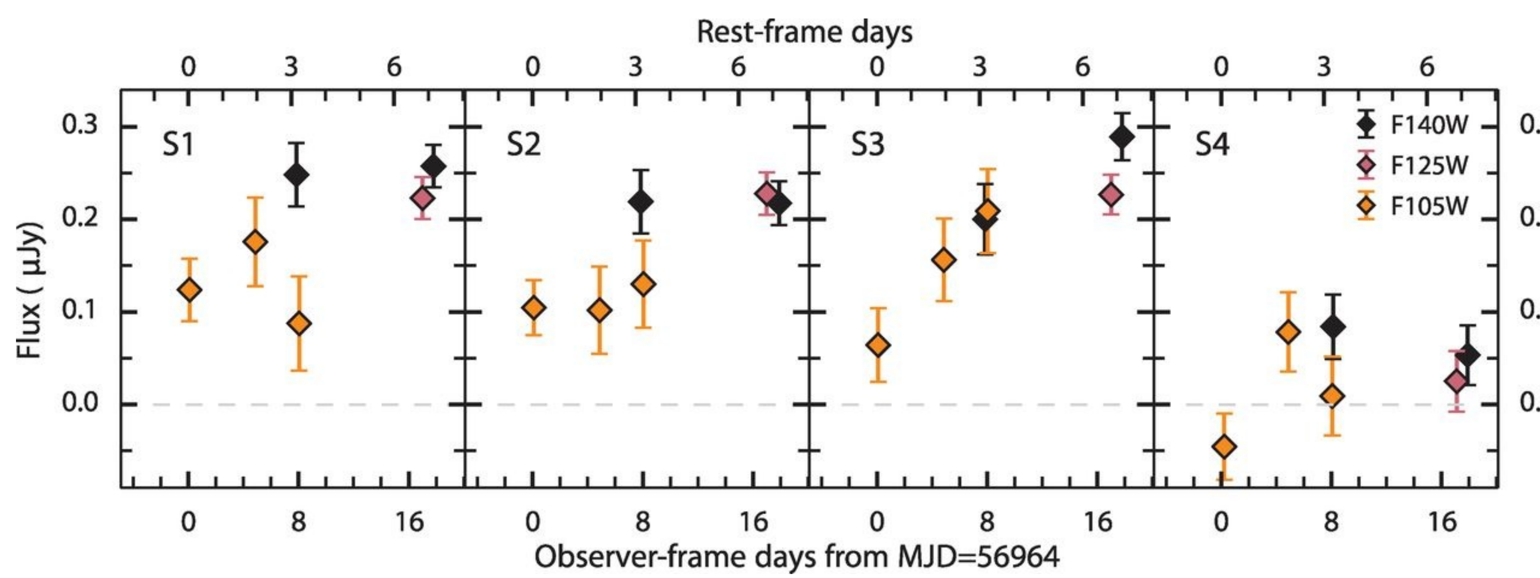
# GP Fit with optimised hyperparameters



# Case Study: SN Refsdal

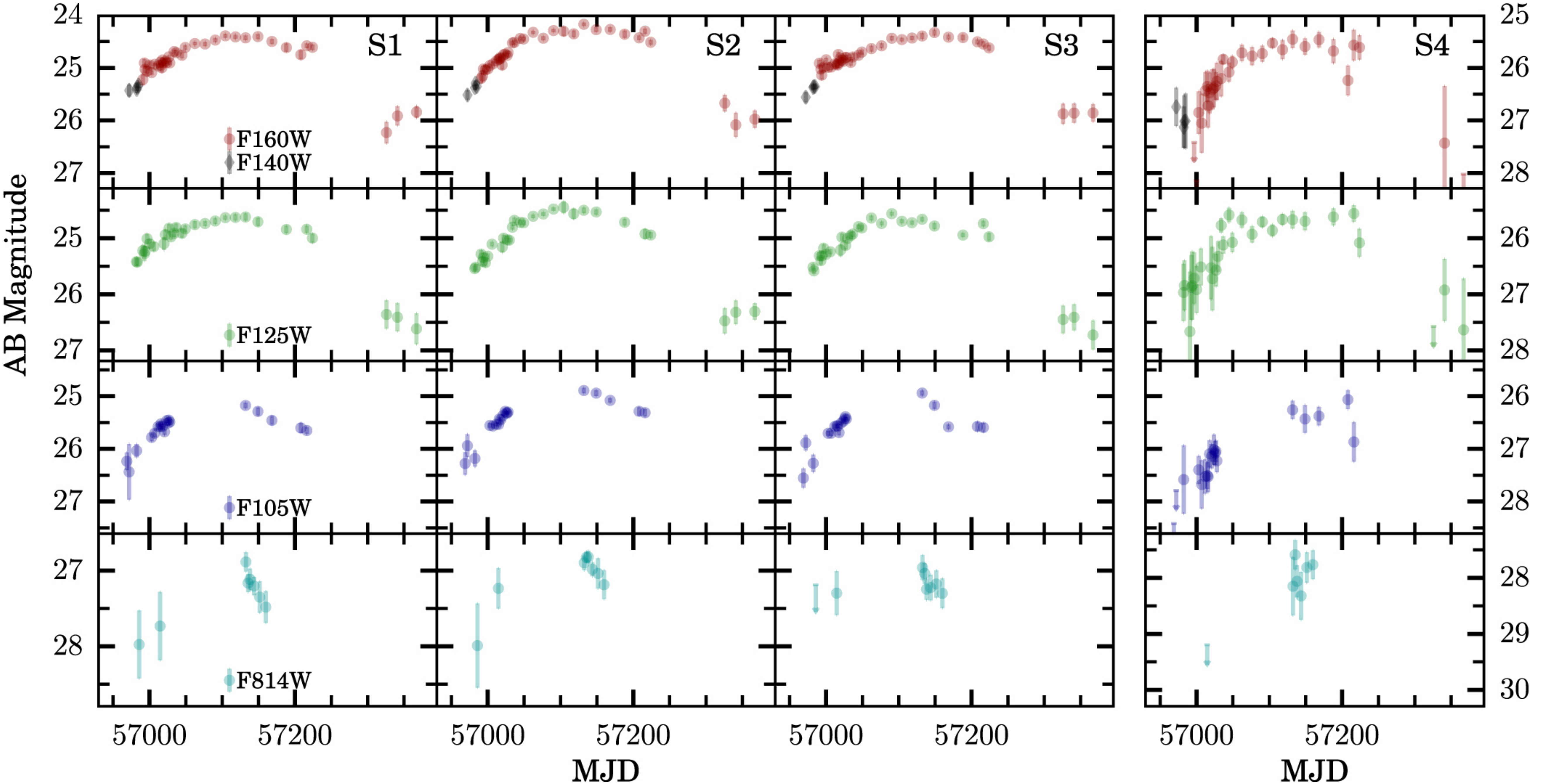


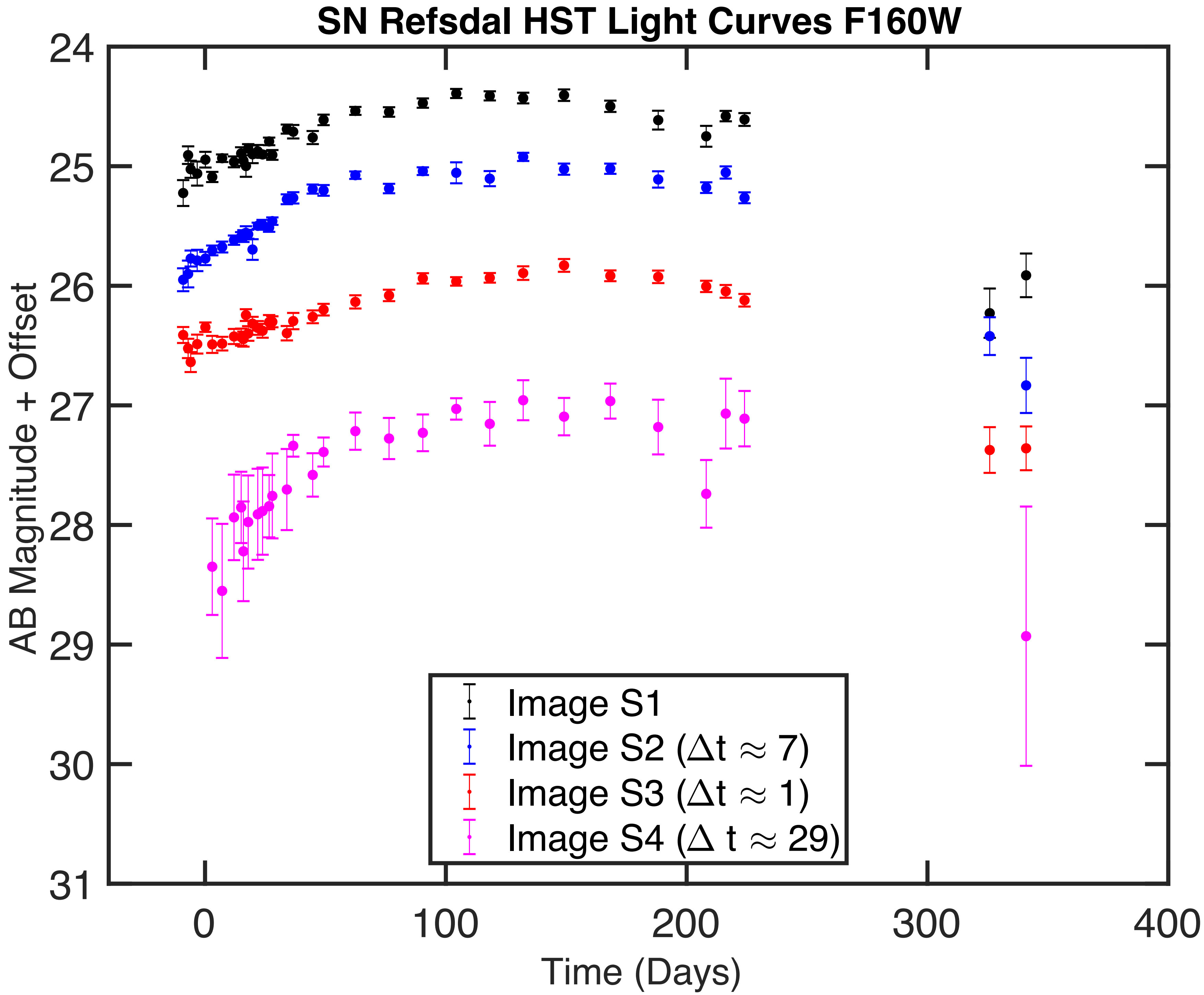

Time Series of SN brightnesses of each image: S1-S4

# 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 $\lambda \approx 1.6$ μm




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

# Construct Bayesian Model

$$f(t) \sim \mathcal{GP}[c, k(t, t,')] \qquad k(t, t') = A^2 \exp[-(t - t')^2/\tau^2]$$

For Image S1 with observations $y_1$ at times $t_1$

$$\boldsymbol{y}_1(\boldsymbol{t}_1) = \boldsymbol{f}(\boldsymbol{t}_1) + \boldsymbol{\epsilon}_1 \qquad \epsilon_{1,j} \sim N(0, \sigma^2_{1,j})$$

$$j \text{ indexes the observations at times } t_{i,j}$$

Images S#i = 2, 3, 4 with observations $y_i$ at times $t_i$ and (unknown) time delays and magnitude shifts (relative to S1) $(\Delta t_i, \Delta m_i)$

$$\boldsymbol{y}_i(\boldsymbol{t}_i) = \Delta m_i + \boldsymbol{f}(\boldsymbol{t}_i - \Delta t_i) + \boldsymbol{\epsilon}_i \qquad \epsilon_{i,j} \sim N(0, \sigma^2_{i,j})$$

# Construct Bayesian Model

Suppose we knew $(\boldsymbol{\Delta t}, \boldsymbol{\Delta m})$

$$\boldsymbol{Y} \equiv \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 - \Delta m_2 \\ \boldsymbol{y}_3 - \Delta m_3 \\ \boldsymbol{y}_4 - \Delta m_4 \end{pmatrix} \qquad \boldsymbol{T} \equiv \begin{pmatrix} \boldsymbol{t}_1 \\ \boldsymbol{t}_2 - \Delta t_2 \\ \boldsymbol{t}_3 - \Delta t_3 \\ \boldsymbol{t}_4 - \Delta t_4 \end{pmatrix}$$

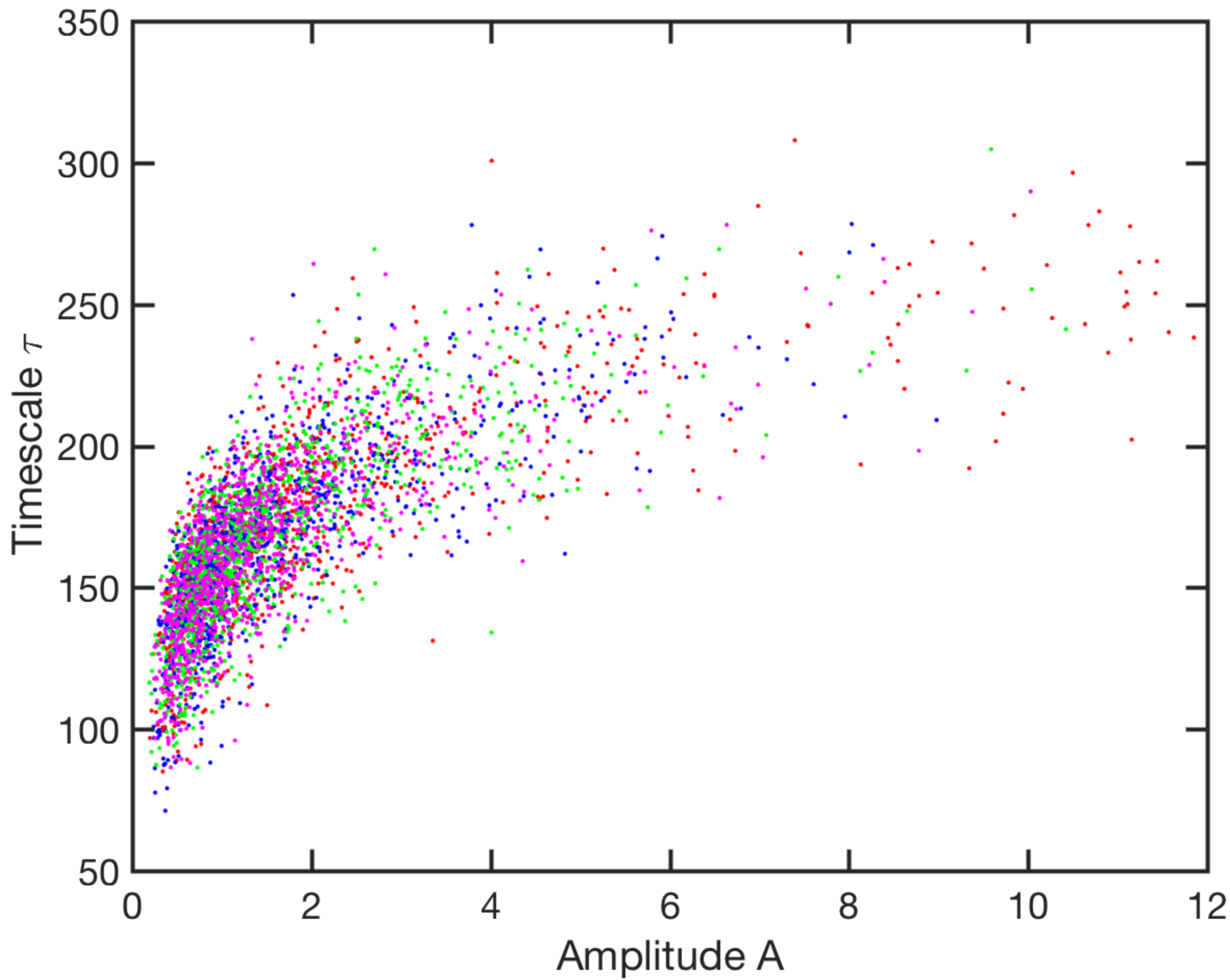
$$P(\boldsymbol{Y} \,|\, \boldsymbol{\Delta m}, \boldsymbol{\Delta t}, A, \tau) = N(\boldsymbol{Y} | \mathbf{1}c, \boldsymbol{K}(\boldsymbol{T}, \boldsymbol{T}) + \boldsymbol{W})$$

$$W_{kl} = \delta_{kl}\, \sigma_k^2$$

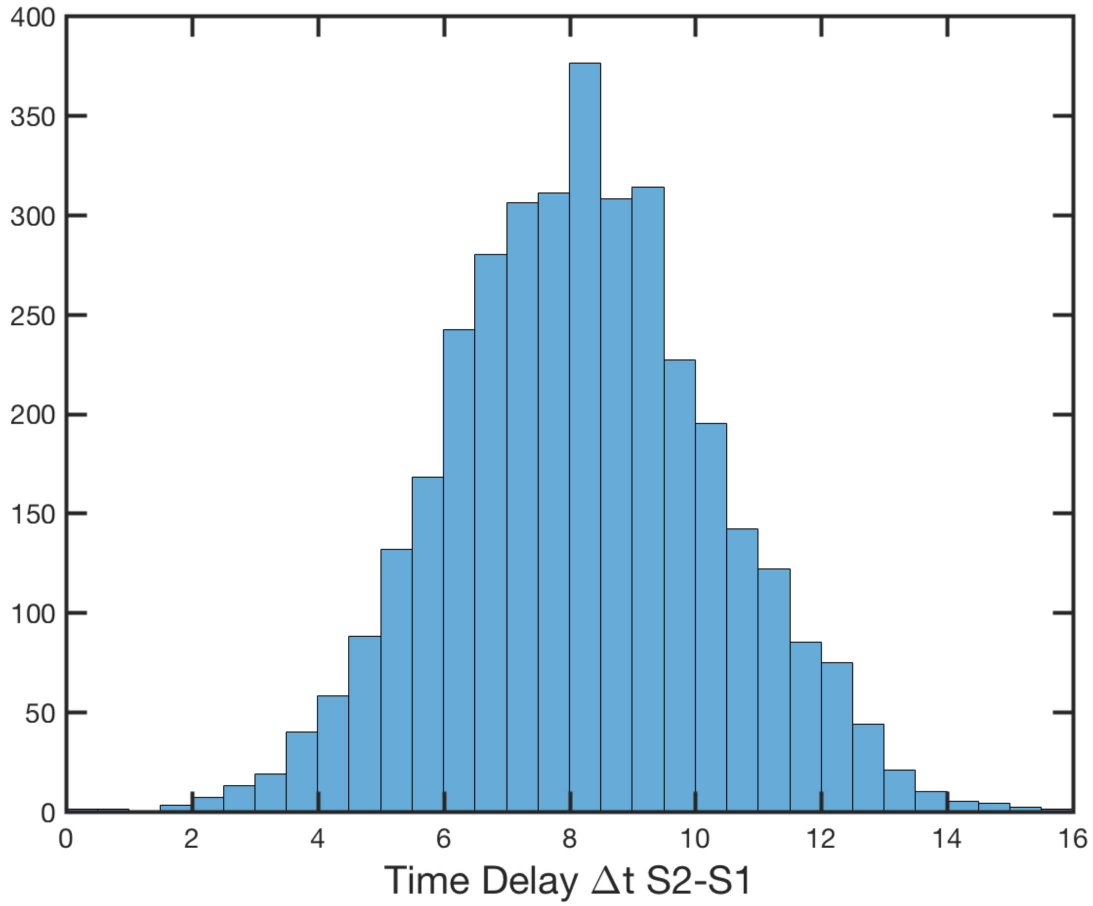
$$\sigma_k^2 = \text{ meas. err. variance for } Y_k$$

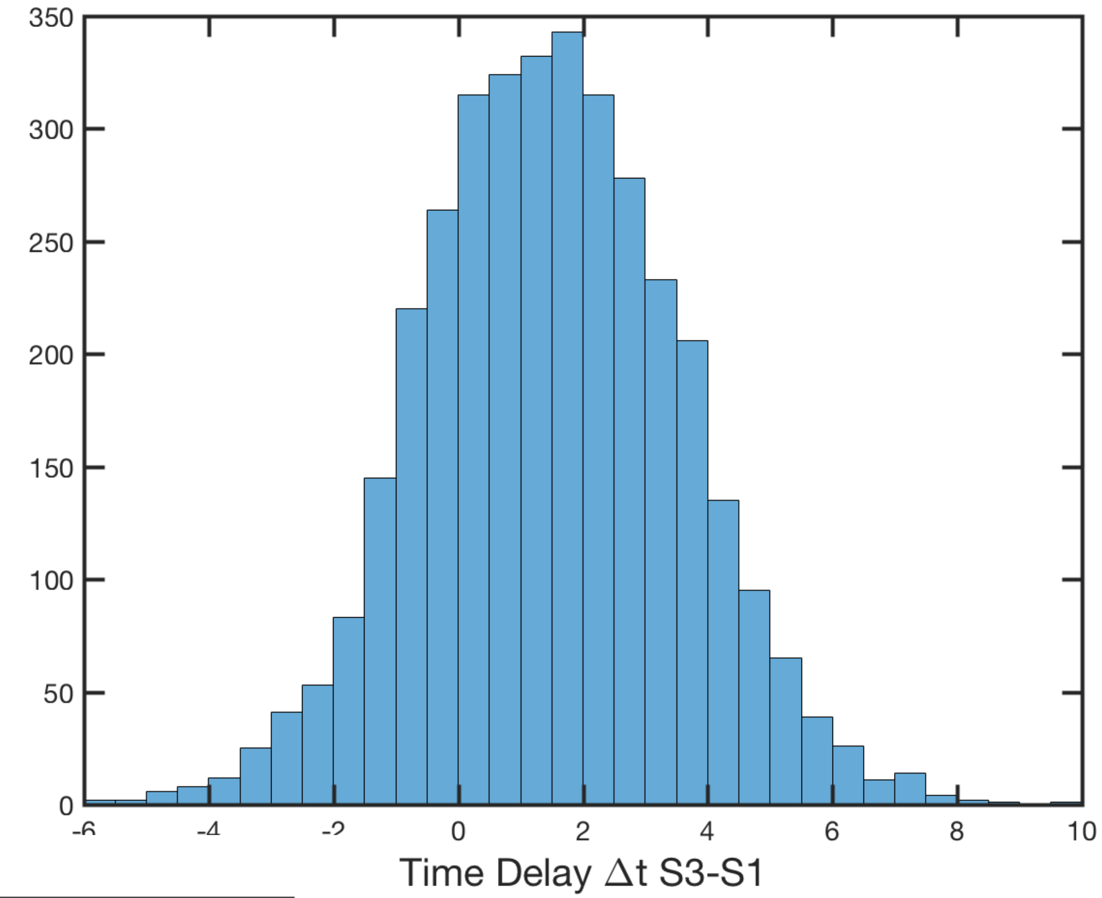
$$P(\boldsymbol{\Delta m}, \boldsymbol{\Delta t}, A, \tau | \boldsymbol{Y}) \propto P(\boldsymbol{Y} \,|\, \boldsymbol{\Delta m}, \boldsymbol{\Delta t}, A, \tau) \times P(\boldsymbol{\Delta m}, \boldsymbol{\Delta t}, A, \tau)$$

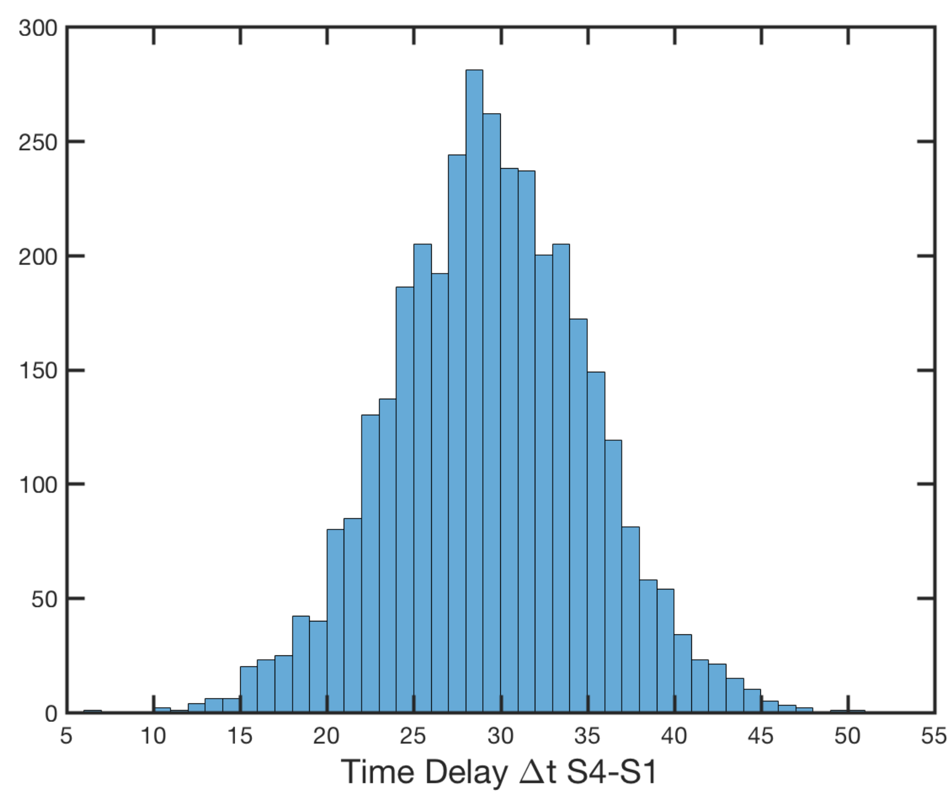
# Metropolis-within-Gibbs results: 8 parameters, 4 chains GP Hyperparameters



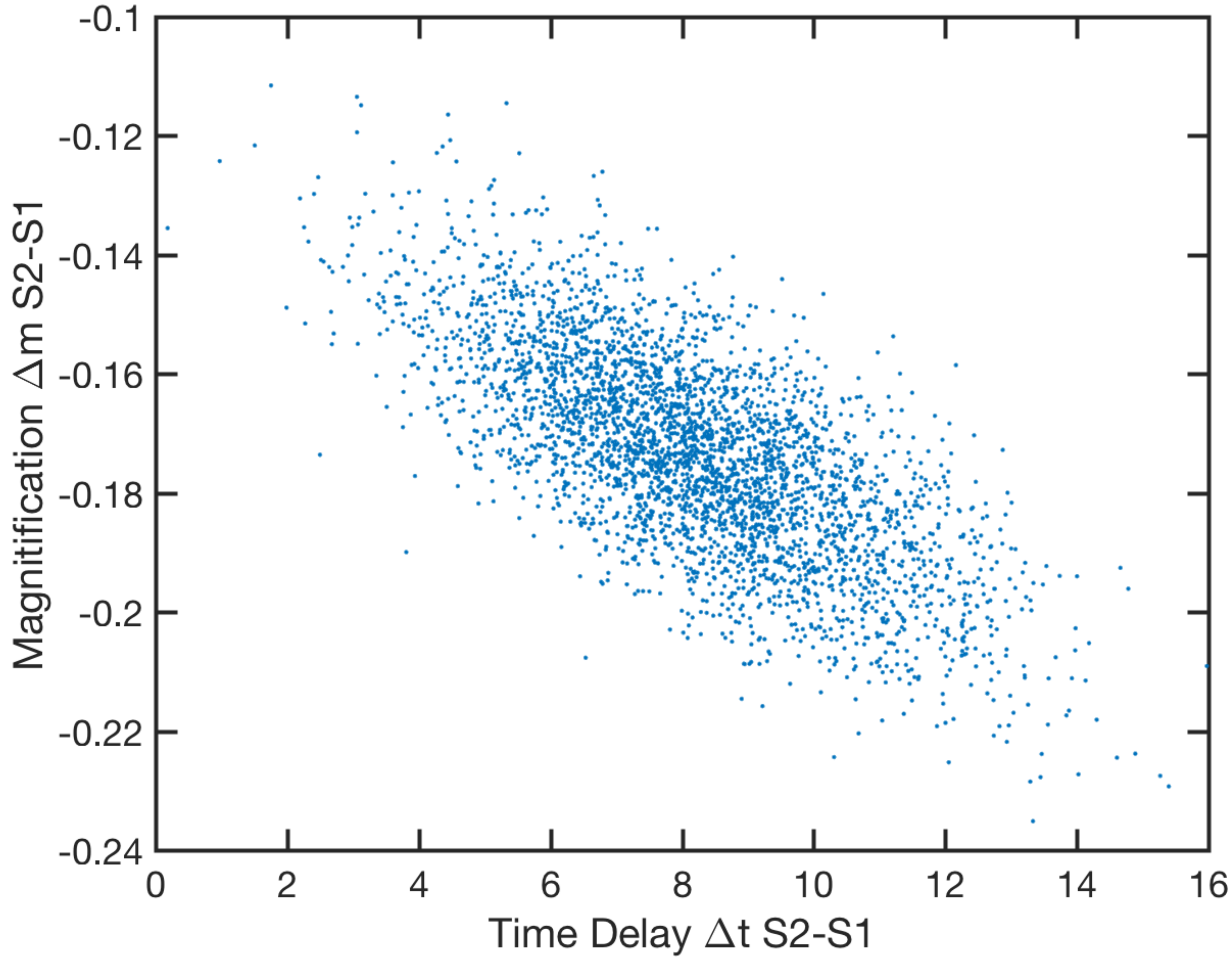
# Metropolis-within-Gibbs results: 4 chains
# Time Delays



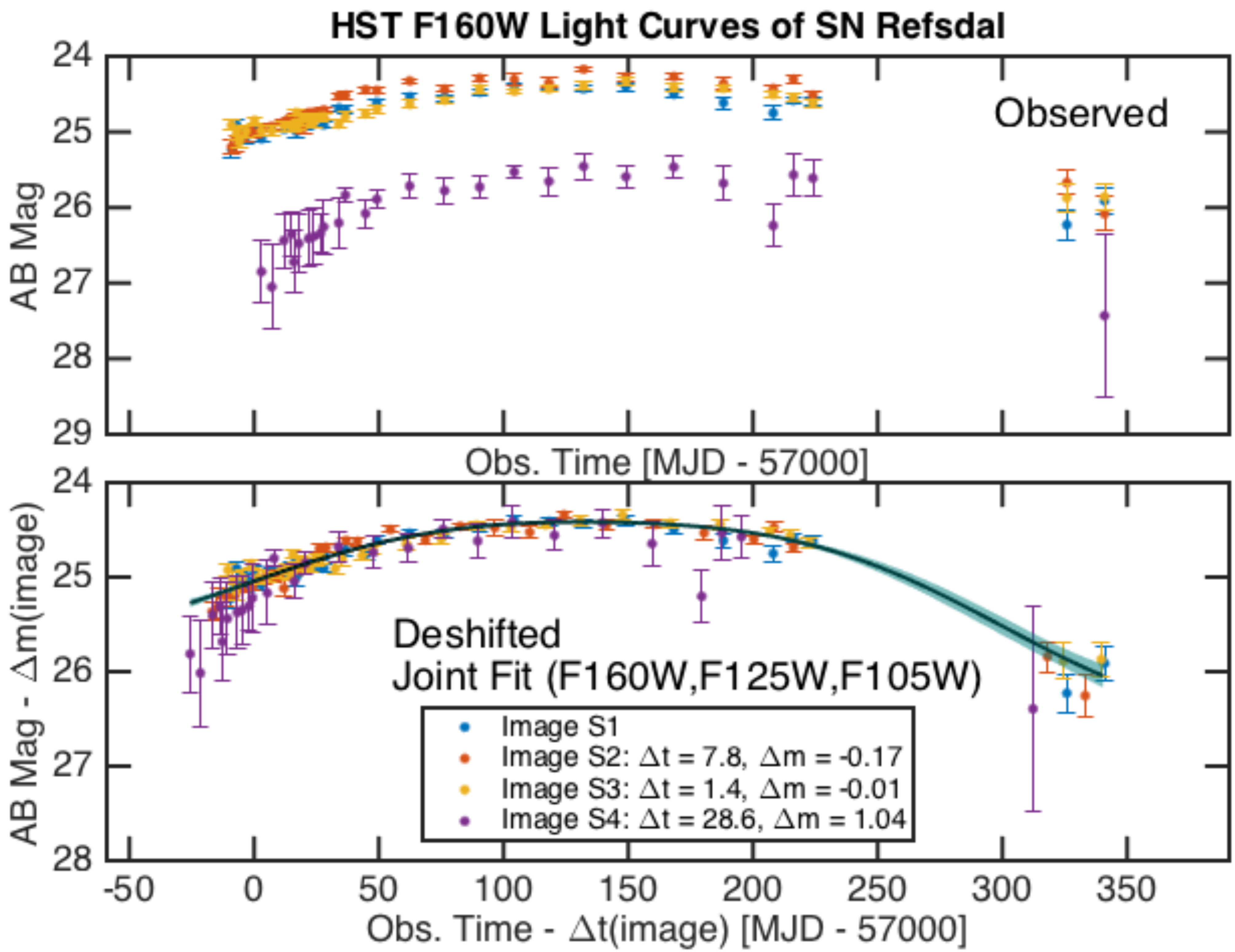



# Metropolis-within-Gibbs results: 4 chains
# Time Delays vs Magnifications



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



# Other covariance functions (R&W Chapter 4)

Squared Exponential gives very smooth curves.

Ornstein Uhlenbeck Process (Damped Random Walk)
Exponential Covariance Function

$$k(t, t') = A^2 \exp(-|t - t'|/\tau)$$

[ Stationary = time translation invariant,
Symmetric: k(t,t') = k(t',t) ]

$$df(t) = \tau^{-1}[\mu - f(t)]dt + \sigma dW_t$$

Mean-Reversion Drag Term        Random Walk

$\tau =$ mean-reversion timescale

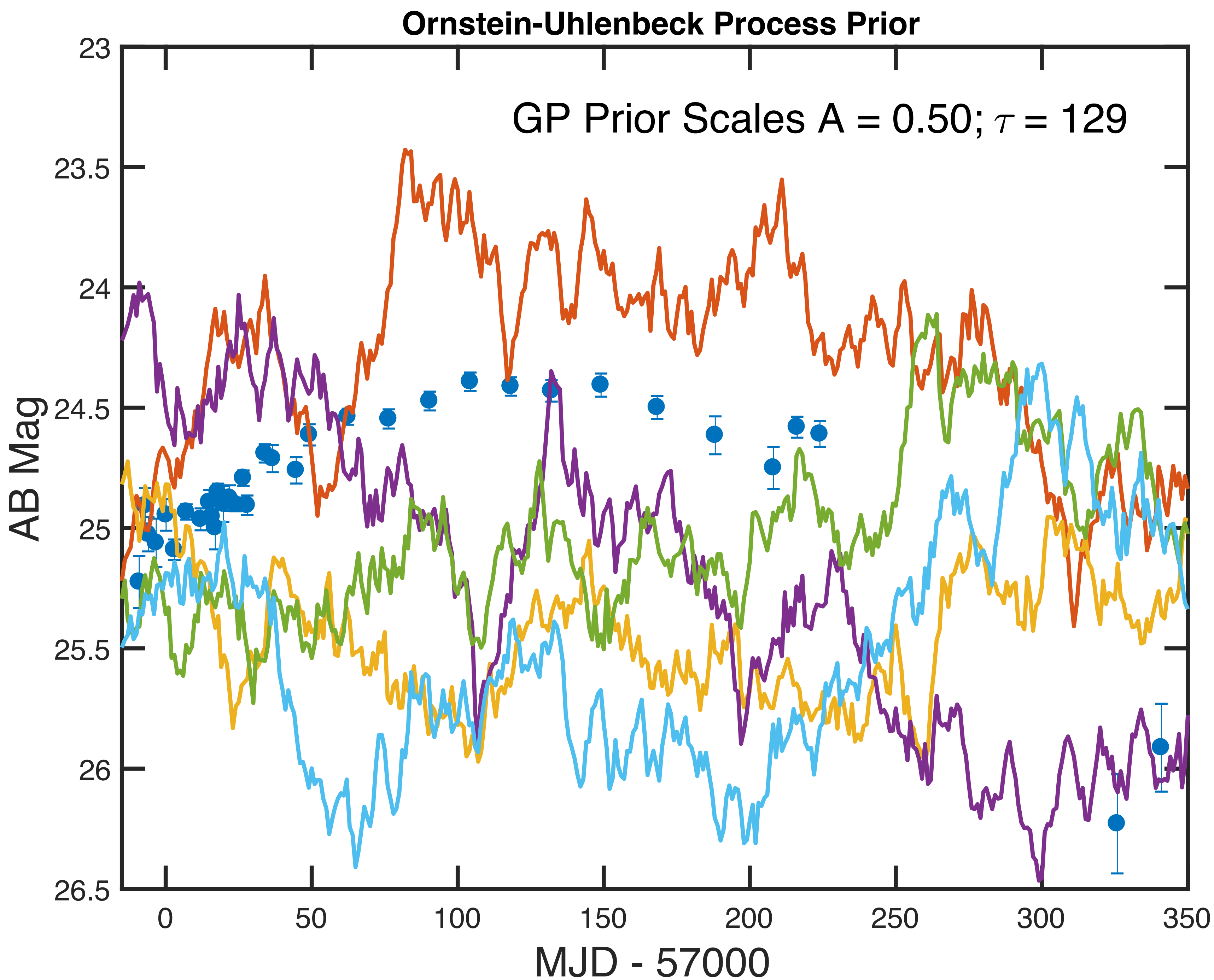
$A^2 = \tau\sigma^2/2$

$\mu =$ long-term mean

$\sigma =$ volatility

# Ornstein Uhlenbeck Process (Damped Random Walk)
# Exponential Covariance Function

$$k(t,t') = A^2 \exp(-|t-t'|/\tau)$$




Everywhere continuous but not differentiable

# Other covariance functions (R&W Chapter 4)

## Ornstein Uhlenbeck Process (Damped Random Walk)
## Exponential Covariance Function

$$k(t,t') = A^2 \exp(-|t-t'|/\tau)$$

## v = 1/2 Special Case of Matern kernel:

**The Matérn Class of Covariance Functions**

The *Matérn class* of covariance functions is given by

$$k_{\mathrm{Matern}}(r) \;=\; \frac{2^{1-\nu}}{\Gamma(\nu)}\Big(\frac{\sqrt{2\nu}r}{\ell}\Big)^{\nu} K_\nu\Big(\frac{\sqrt{2\nu}r}{\ell}\Big), \tag{4.14}$$

with positive parameters $\nu$ and $\ell$, where $K_\nu$ is a modified Bessel function

When v = p + 1/2 is half-integer: Exponential x p-Polynomial

$$k_{\nu=p+1/2}(r) \;=\; \exp\Big(-\frac{\sqrt{2\nu}r}{\ell}\Big)\frac{\Gamma(p+1)}{\Gamma(2p+1)}\sum_{i=0}^{p}\frac{(p+i)!}{i!(p-i)!}\Big(\frac{\sqrt{8\nu}r}{\ell}\Big)^{p-i}$$

# Periodic Covariance Functions

$$k(t, t') = A^2 \exp\left(-\frac{2r^2}{l^2}\sin^2(\pi(t-t')/T)\right)$$

(Example Sheet: Periodic kernel can be derived for any Gaussian process on **R²**)