Astrostatistics: 01 Mar 2019

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

- Example Class 2: Tue, 05 March, 12-2pm MR9
- Example Class 3: Thu, 14 March, 12-2pm MR9
- Example Class 4: Probably ~ April 29-30
- Begin Gaussian Processes in Astrophysics
- Case Study: application to gravitationally lensed quasars & supernovae

MCMC in Practice

1. Find the mode(s), using optimisation.

2. Begin multiple (parallel) chains at starting positions dispersed around the mode(s).

3. Scale Metropolis proposals to tune 25-50% acceptance rate (depending on dimension of jump)

4. Use proposal cov matrix that reflects the shape of the posterior

5. After run, look at chains to check for obvious problems

6. Compute Gelman-Rubin ratio comparing within-chain-variance to between-chain variance to check that chains are well-mixed (should be very close to 1), and assess burn-in

7. Compute autocorrelation timescale and effective sample size to make sure you have enough *independent* samples for inference.

8. If all checks out, remove burn-in, (thin), and combine chains to compute posterior inferences

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: $\boldsymbol{\Sigma}$ 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 $\mathbf{Var}[\mathbf{f}] = \mathbf{Cov}[\mathbf{f}, \mathbf{f}^{\mathbf{T}}]$.

- 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:

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

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)

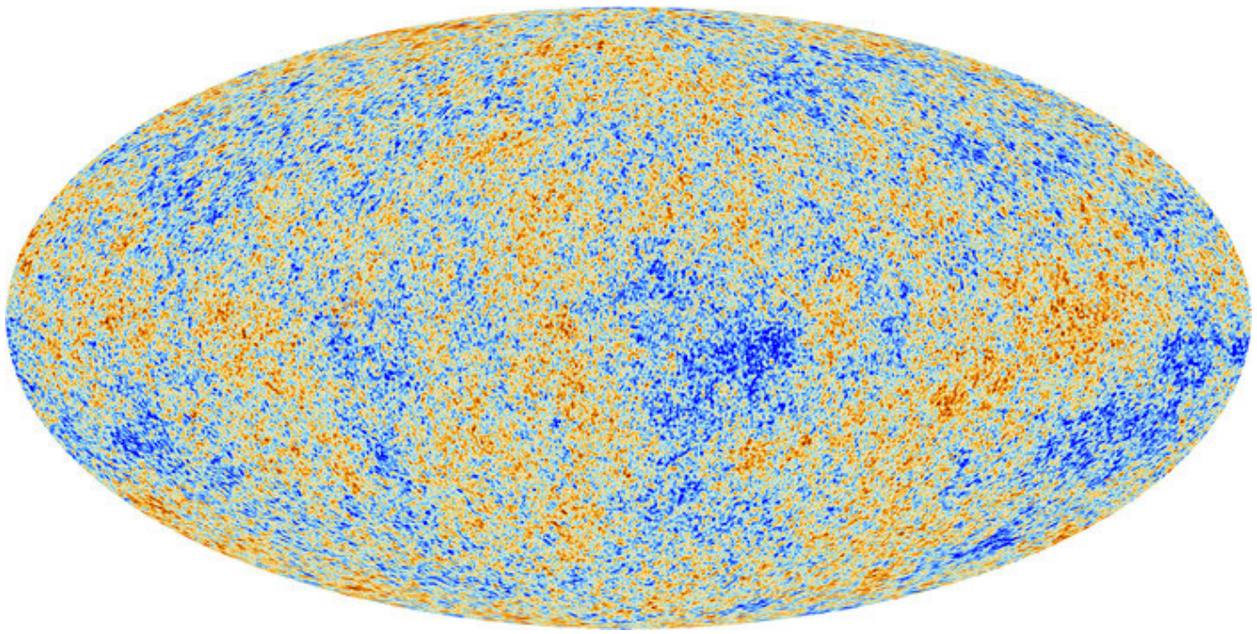
- “Nonparametric” models: flexible functions to use when an accurate astrophysical function is not available or is imperfect

- Nonparametric ~ number of parameters grows with the dataset

- Interpolation/Emulation: To generate a smooth curve going through some observation or simulation points

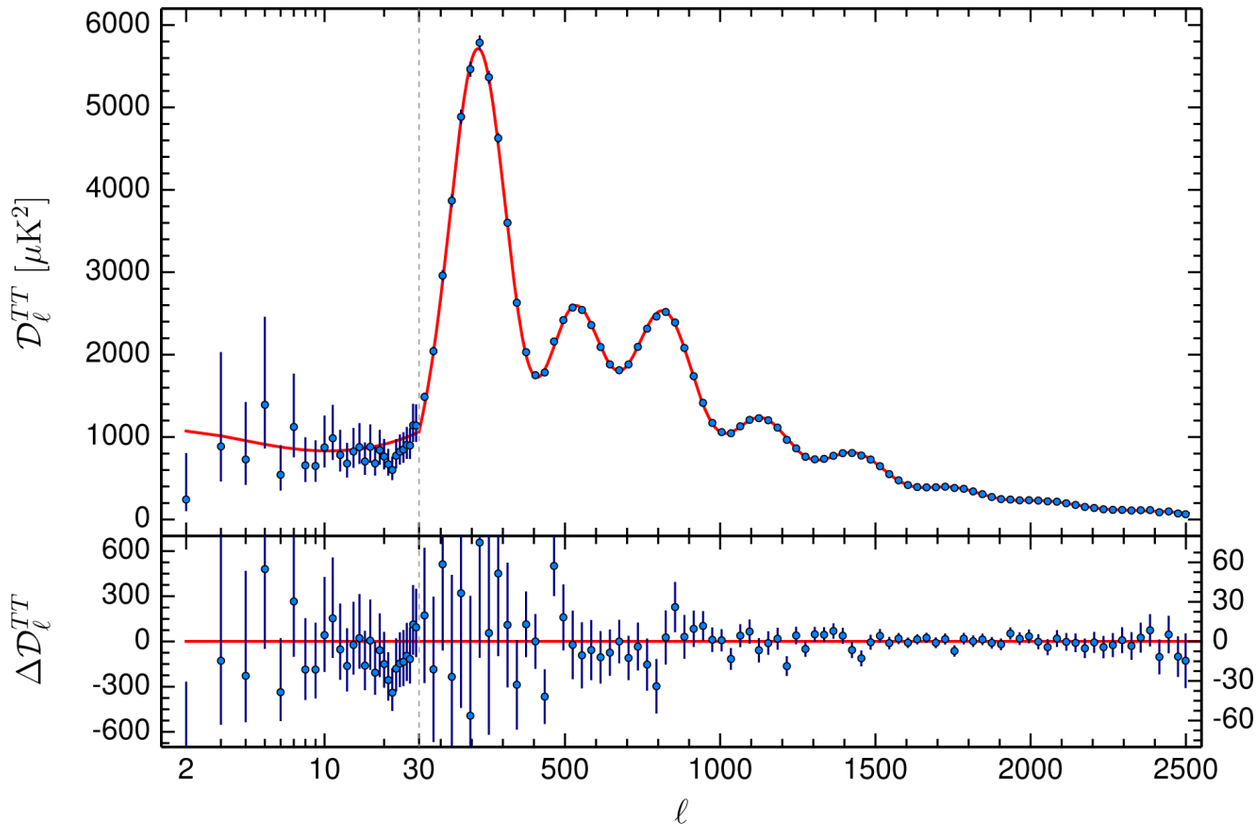
- Correlated noise/error model: When you marginalise out the latent error function, you are effectively accounting for correlated noise over time/space/wavelength.
  e.g. *D. Jones et al.* “Improving Exoplanet Detection Power: Multivariate Gaussian Process Models for Stellar Activity.” arXiv:1711.01318

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

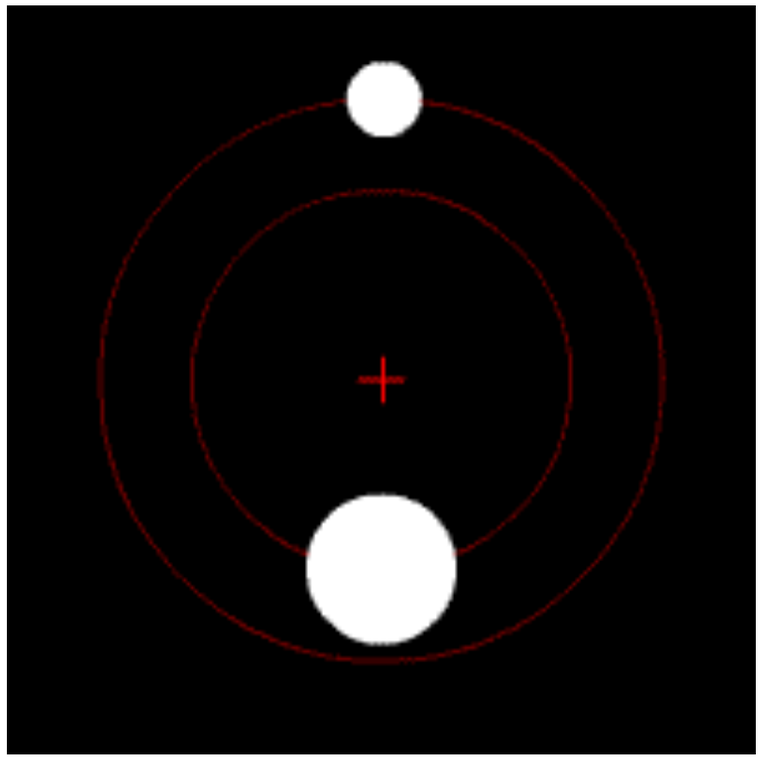


Cosmic Microwave Background (Planck) ~ Gaussian Random Field (mean = 2.7 K, std dev ~ 10^{-5})

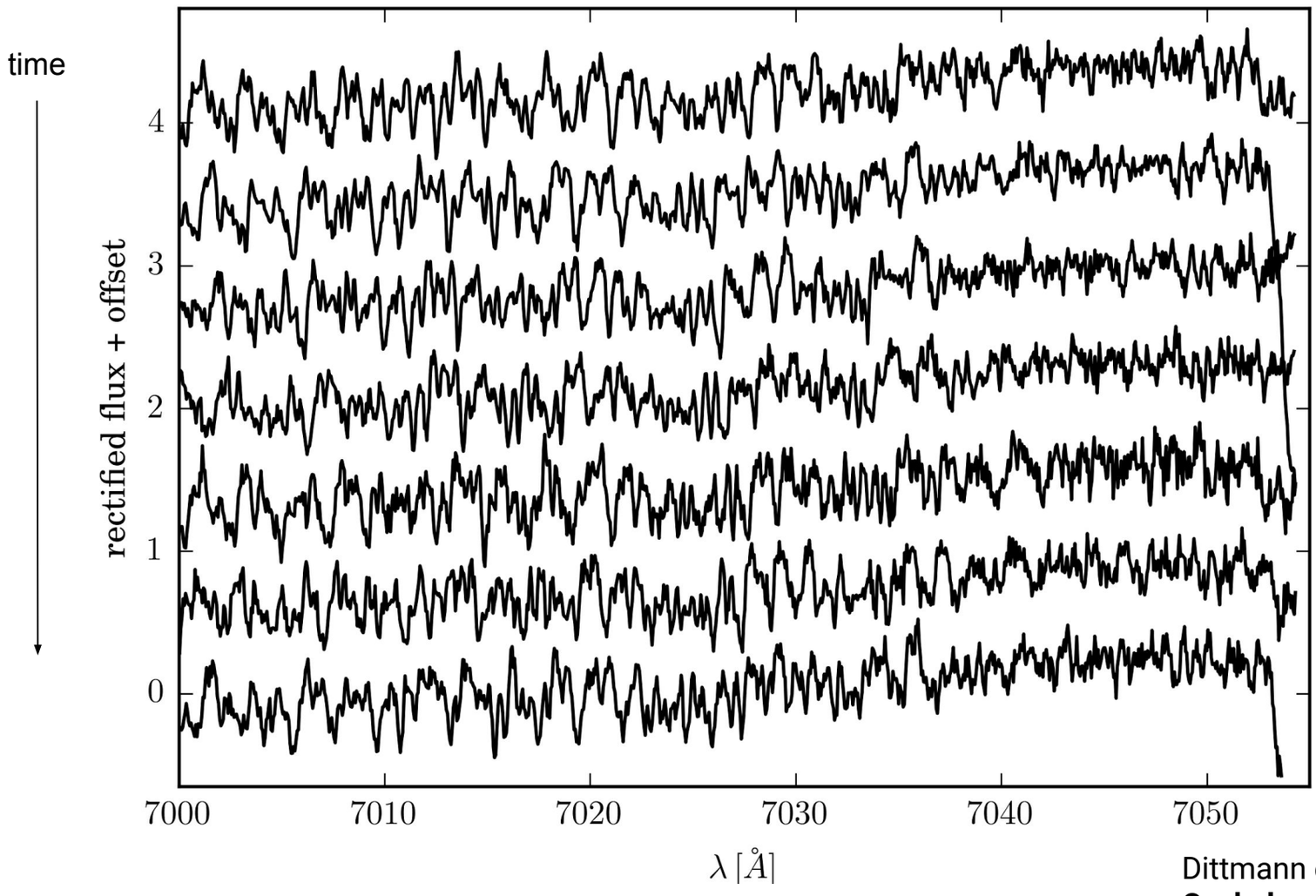
Power Spectrum (~Fourier Transform of Covariance Function)



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



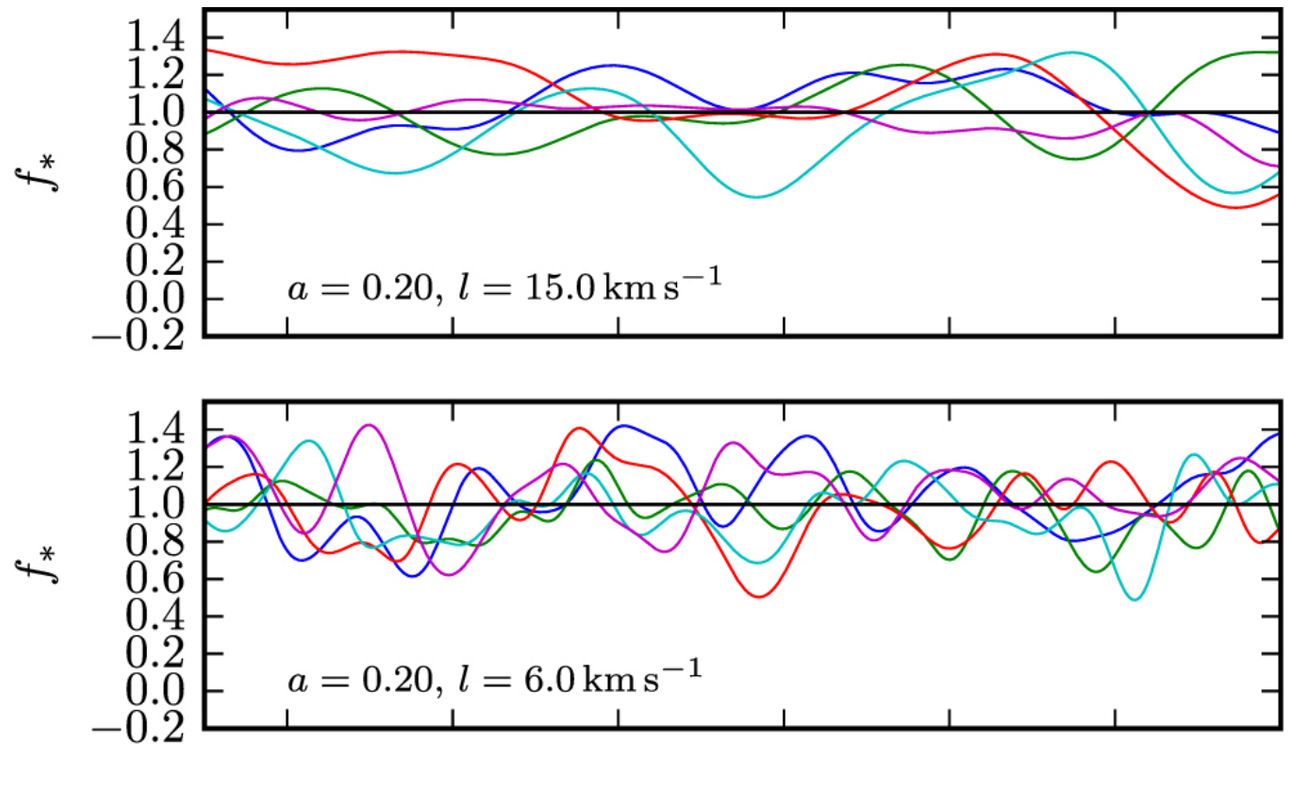
Raw Observations of the LP661-13 M4 Binary



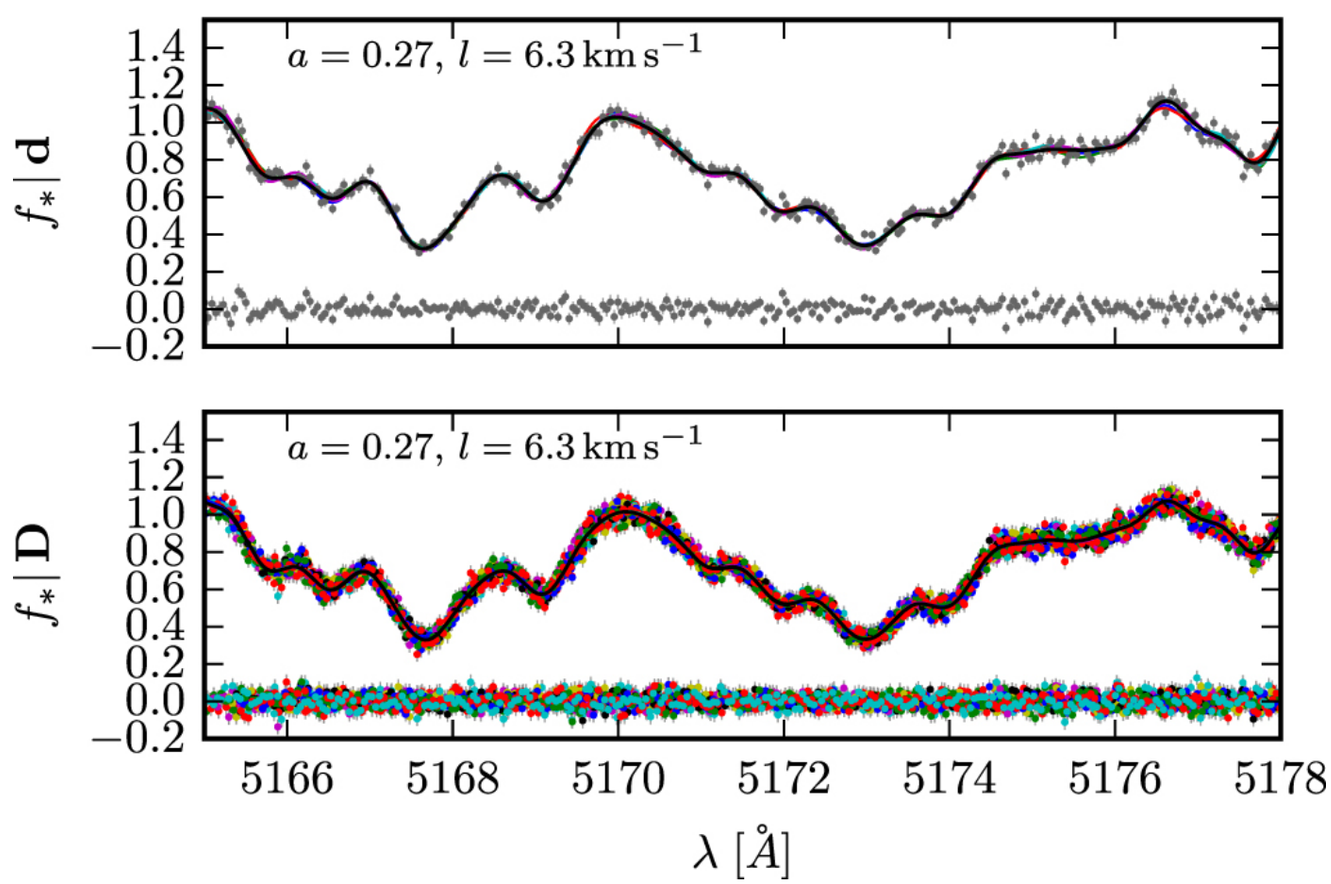
Dittmann et al. 17
Czekala et al. 17a

Example: Gaussian Process: Priors & Posteriors

GP prior (long/short correlation length scales)



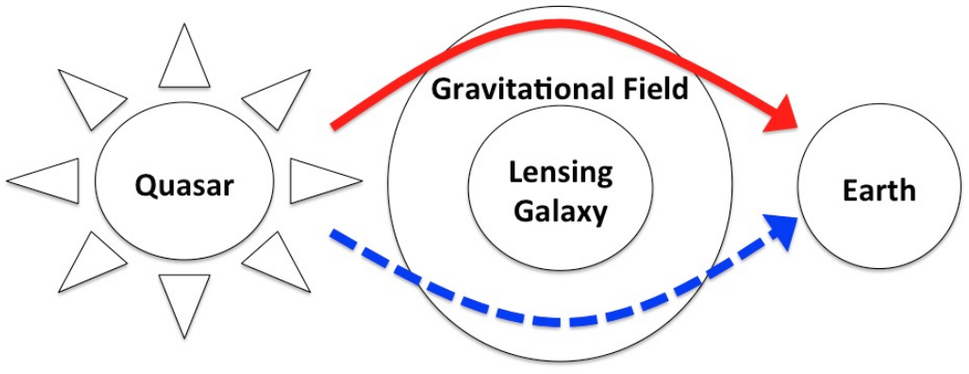
GP Posterior (conditioned on data spectrum d) Inference of latent spectrum

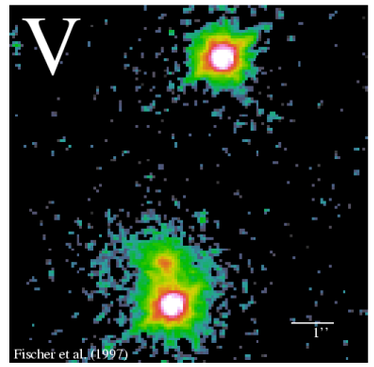


Astrostatistics Case Study:

Bayesian Estimates of Astronomical Time Delays Between Gravitationally Lensed Stochastic Light Curves

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





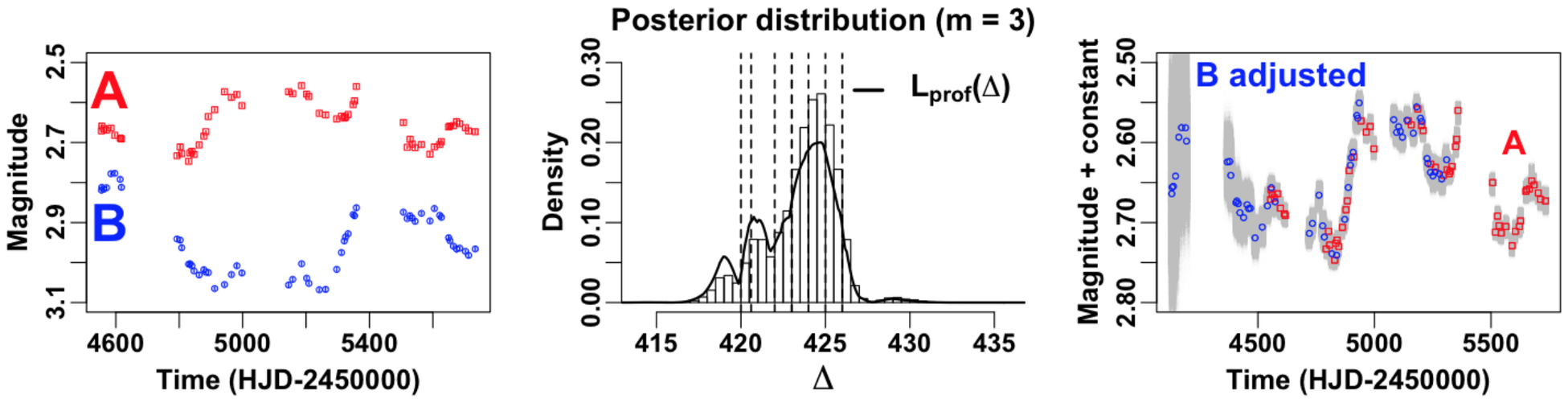
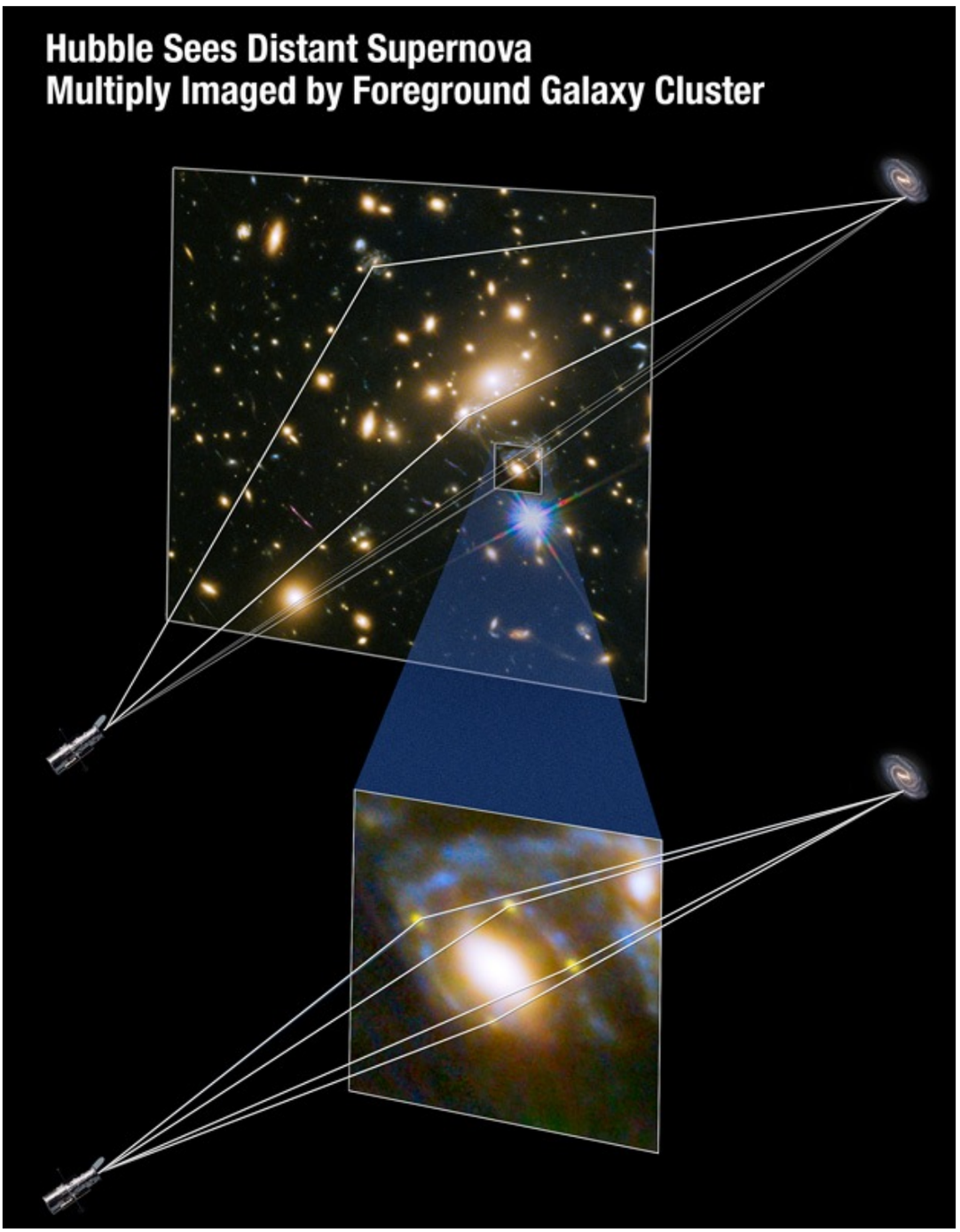
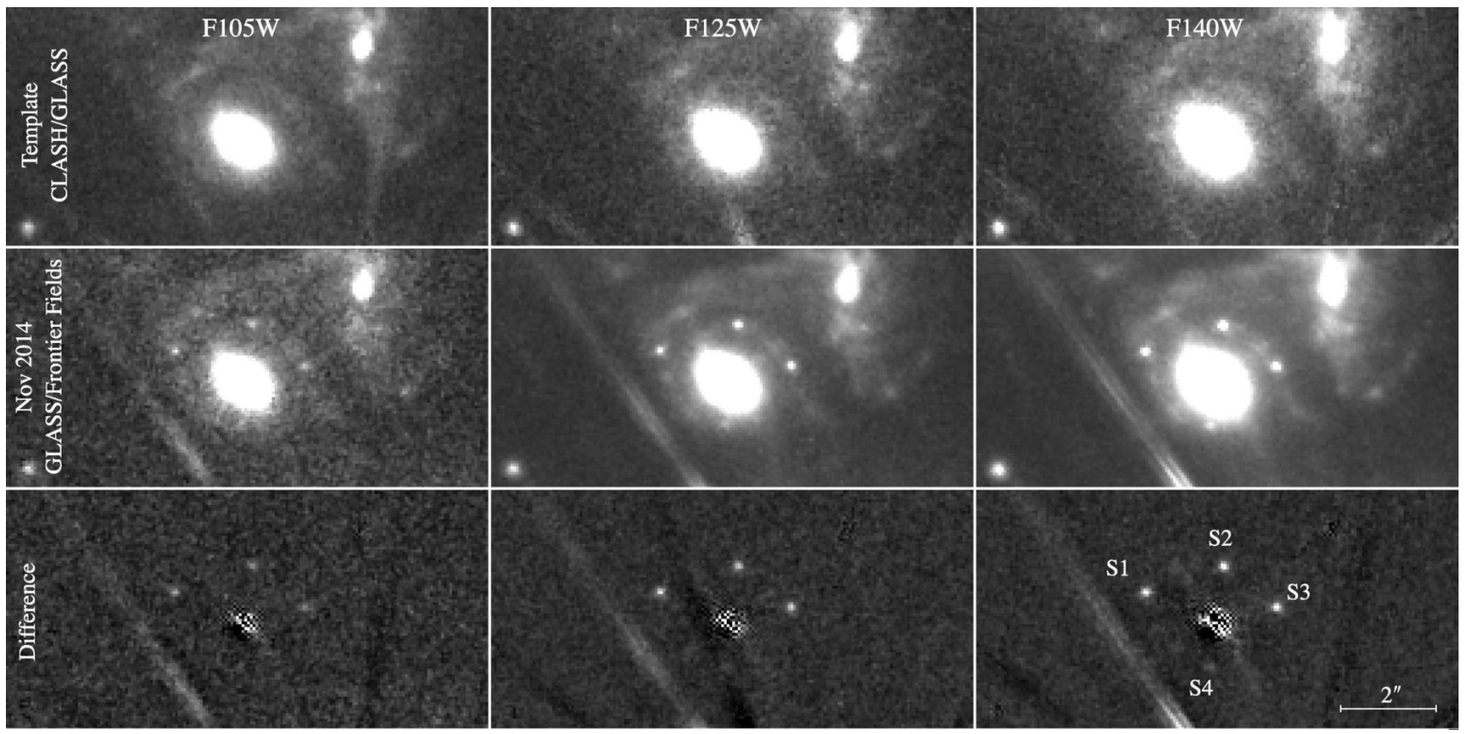


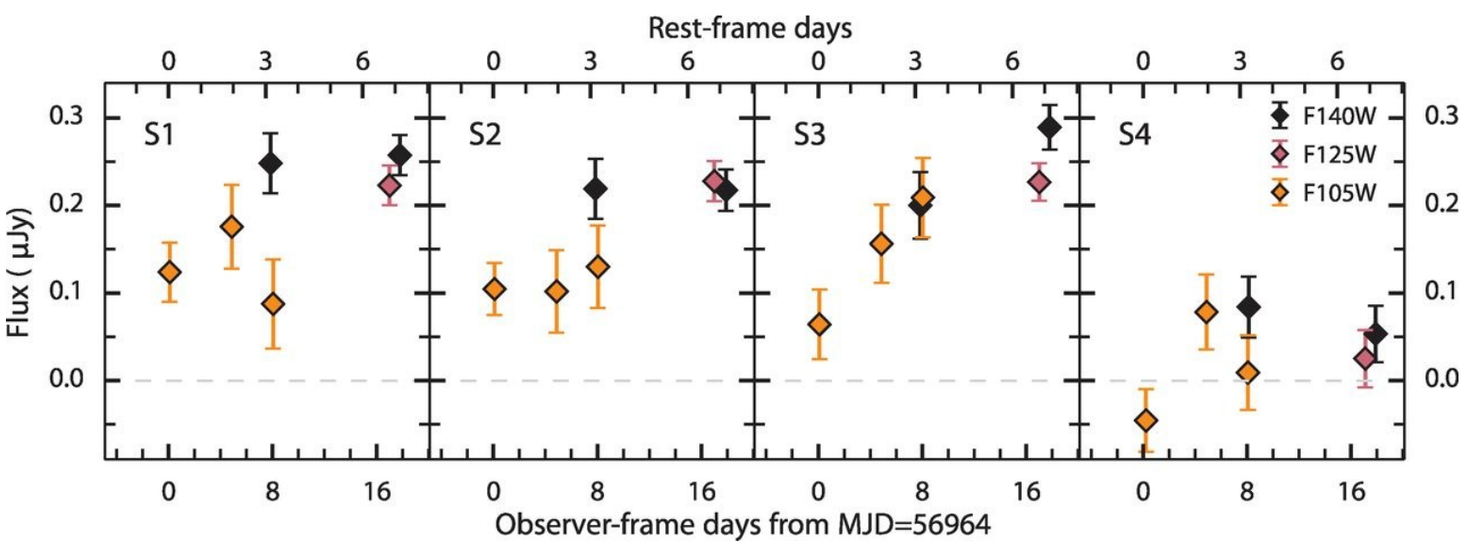
FIG 13. *Observations of Quasar Q0957+561 from Hainline et al. (2012) are plotted in the first panel. The second panel exhibits the marginal posterior distribution of* Δ *with*

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

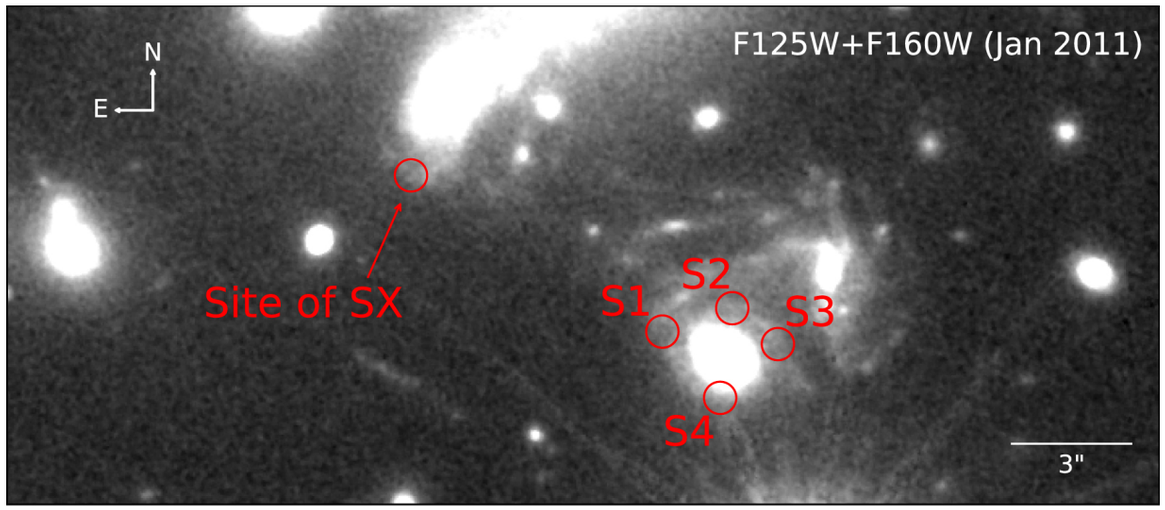


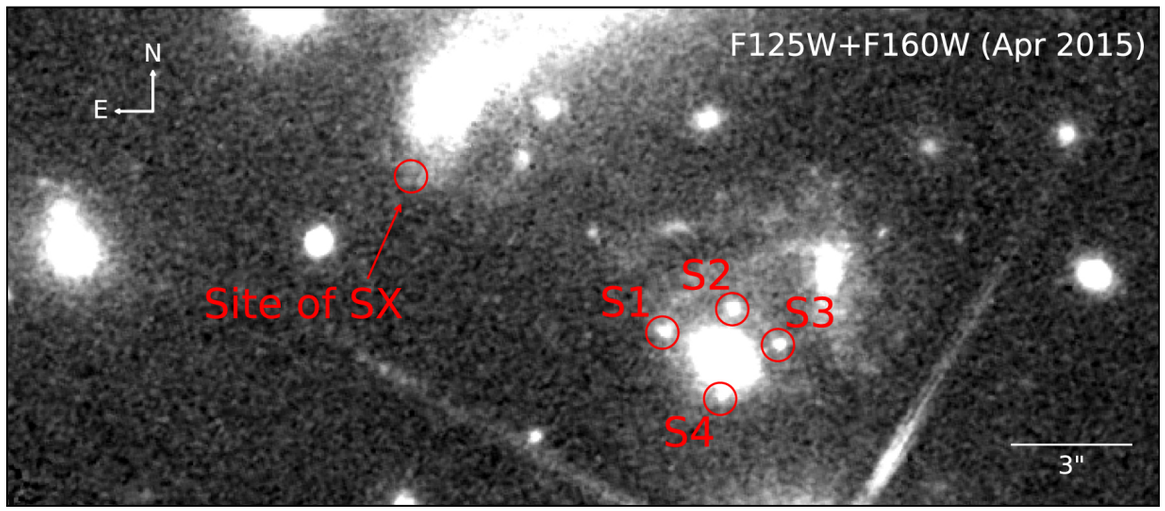


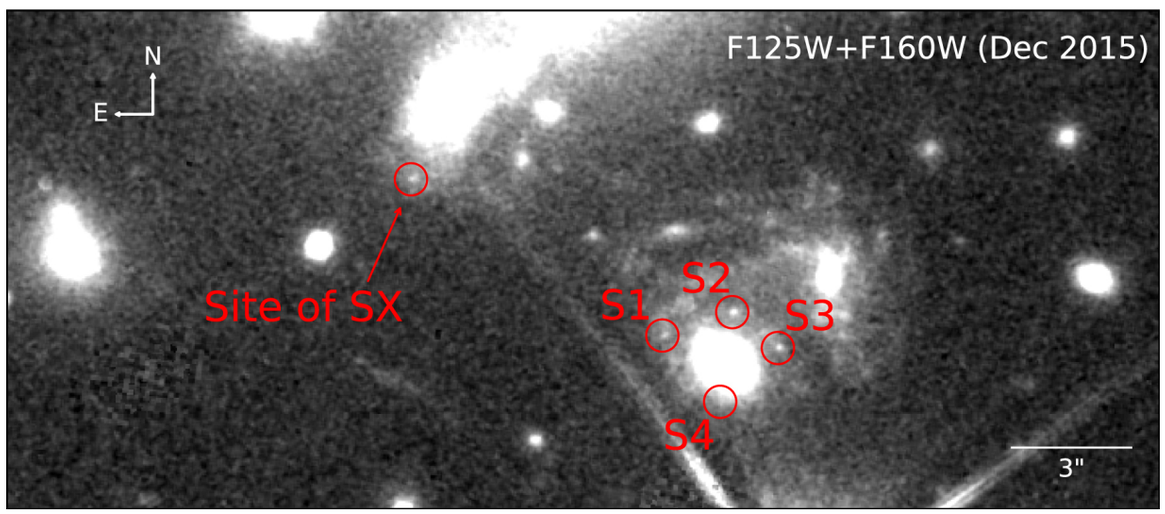


Time Series of SN brightnesses of each image: S1-S4

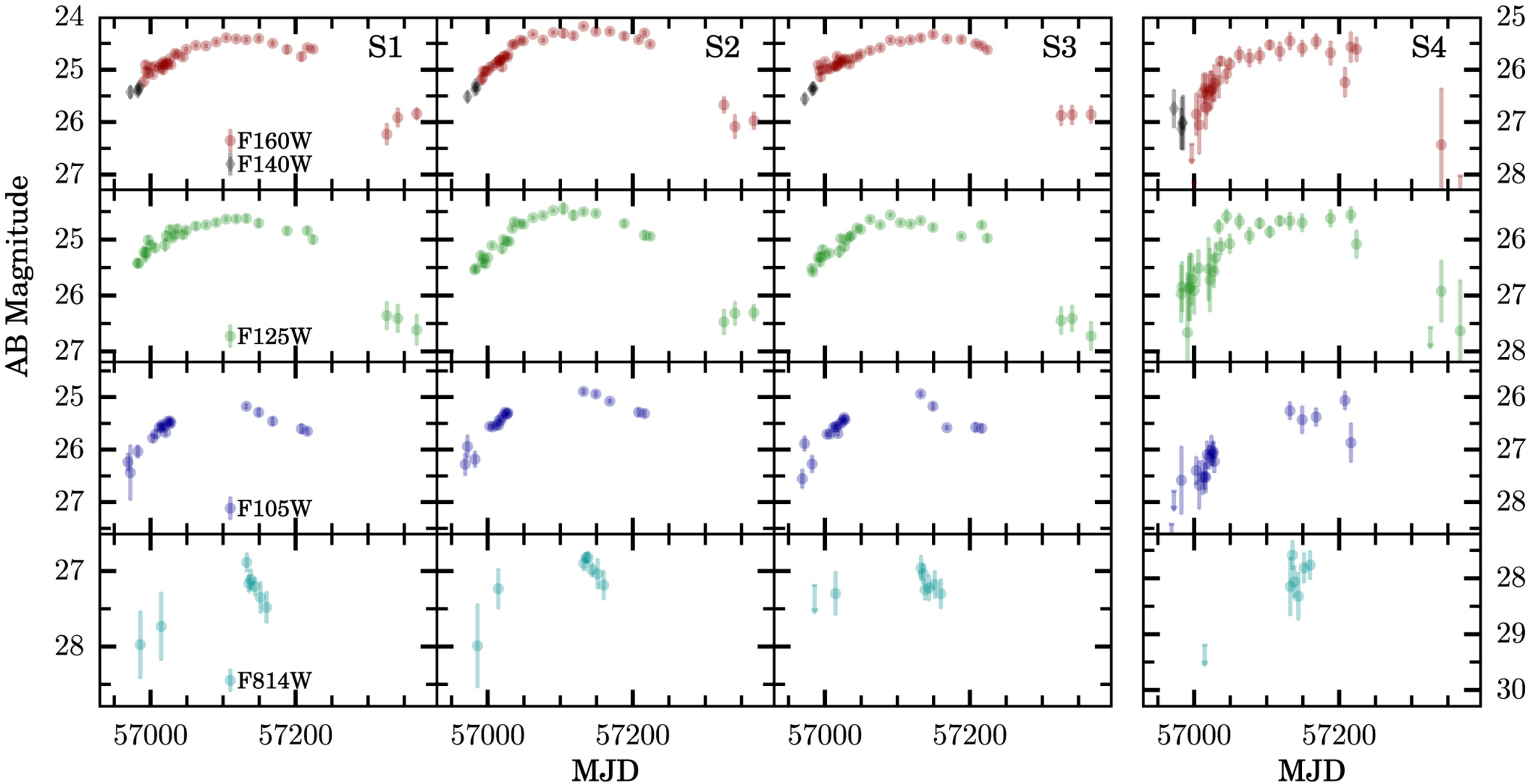
Prediction and Confirmation of the Reappearance of 5th Lensed Image of SN Refsdal (Kelly et al. 2015)





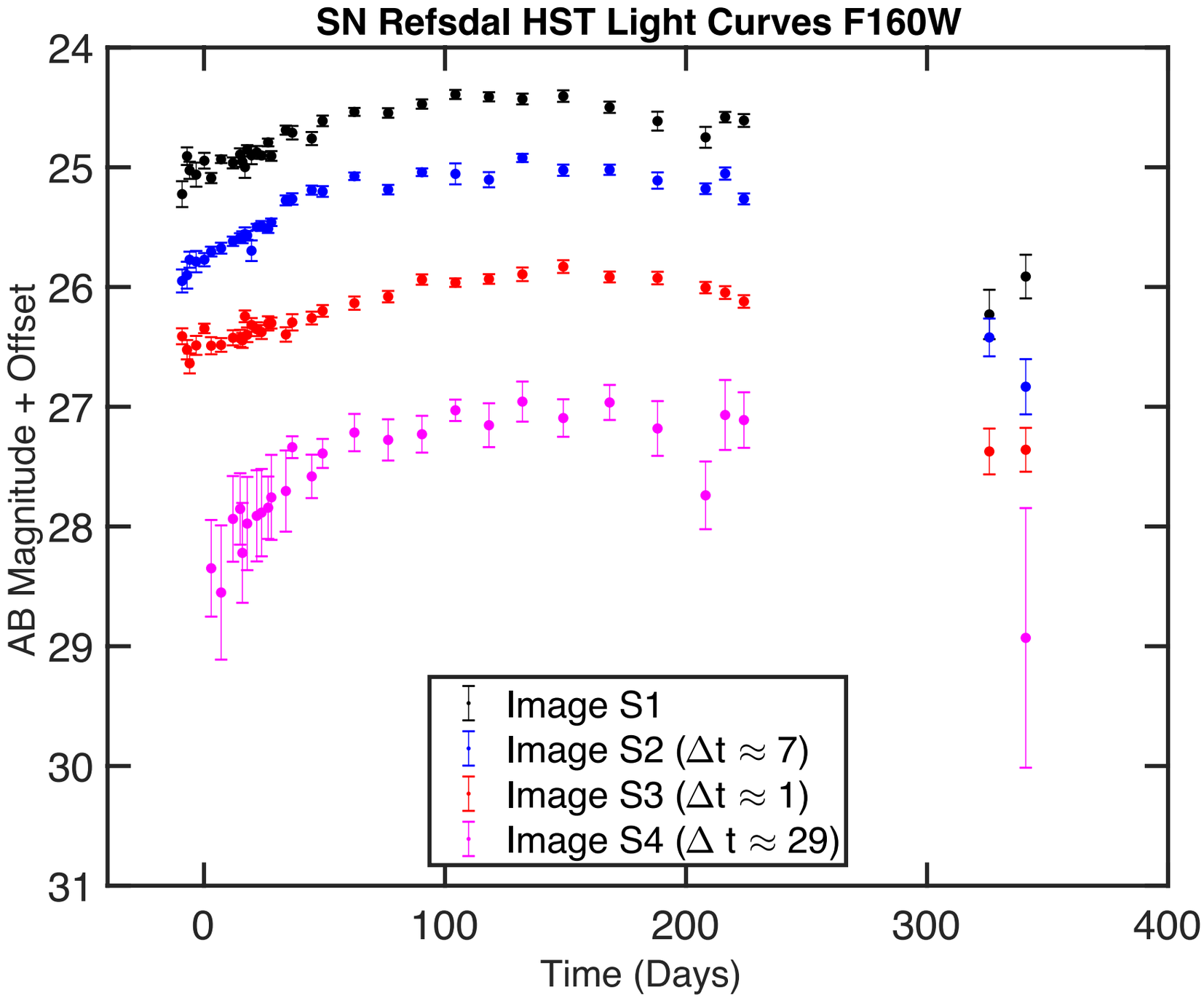


Hubble Space Telescope time series of SN Refsdal multiple images (Rodney et al. 2016)



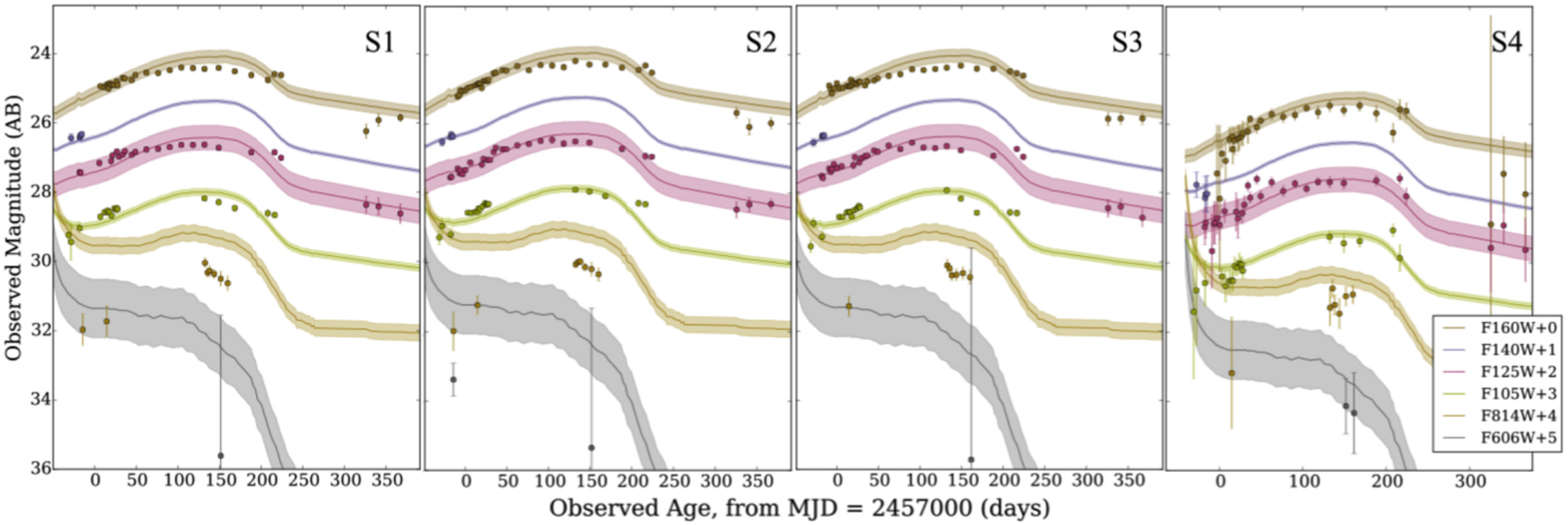
Brightness Time Series [MJD = Modified Julian Day]

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)

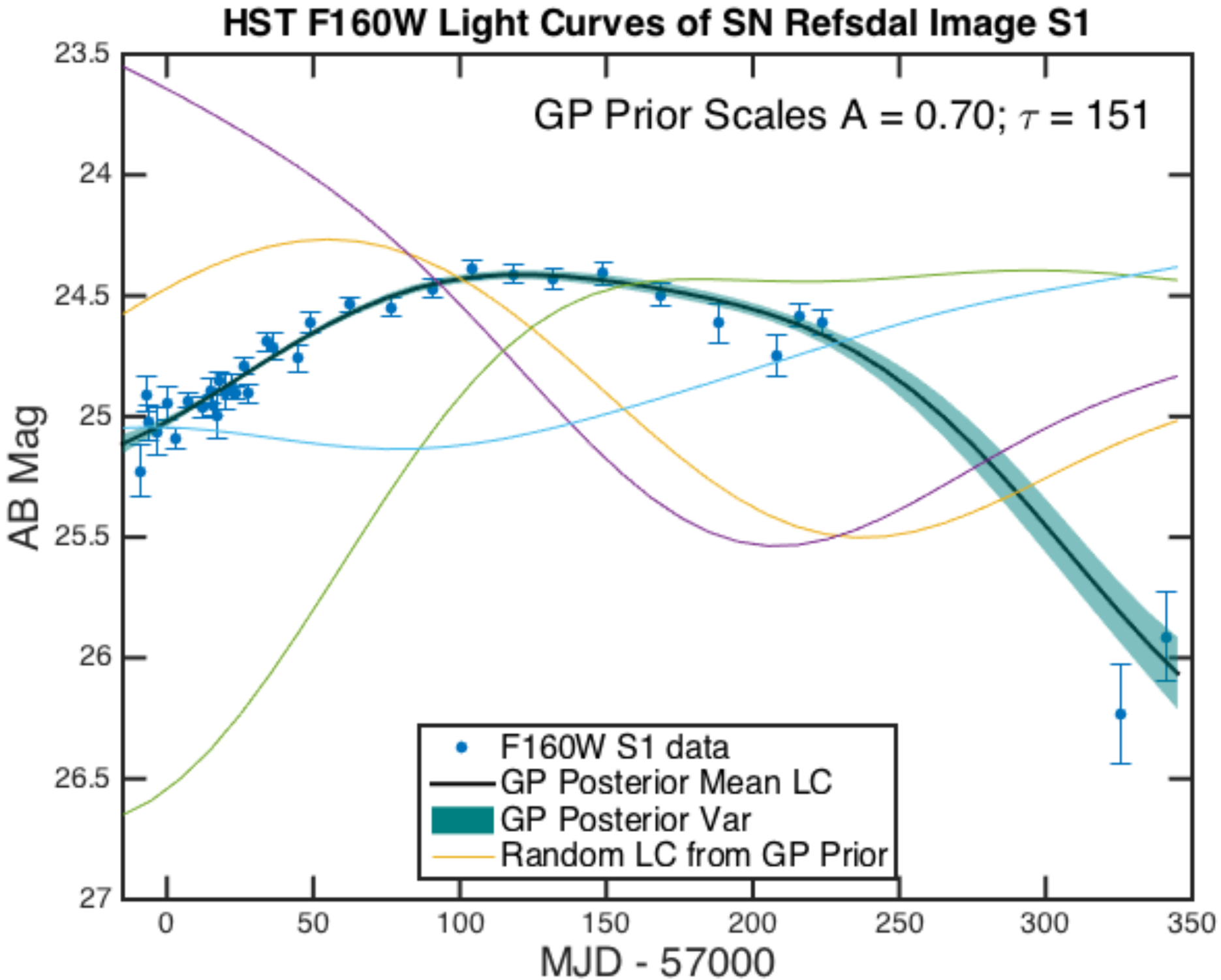


<u>Our Strategy:</u>

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**:

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

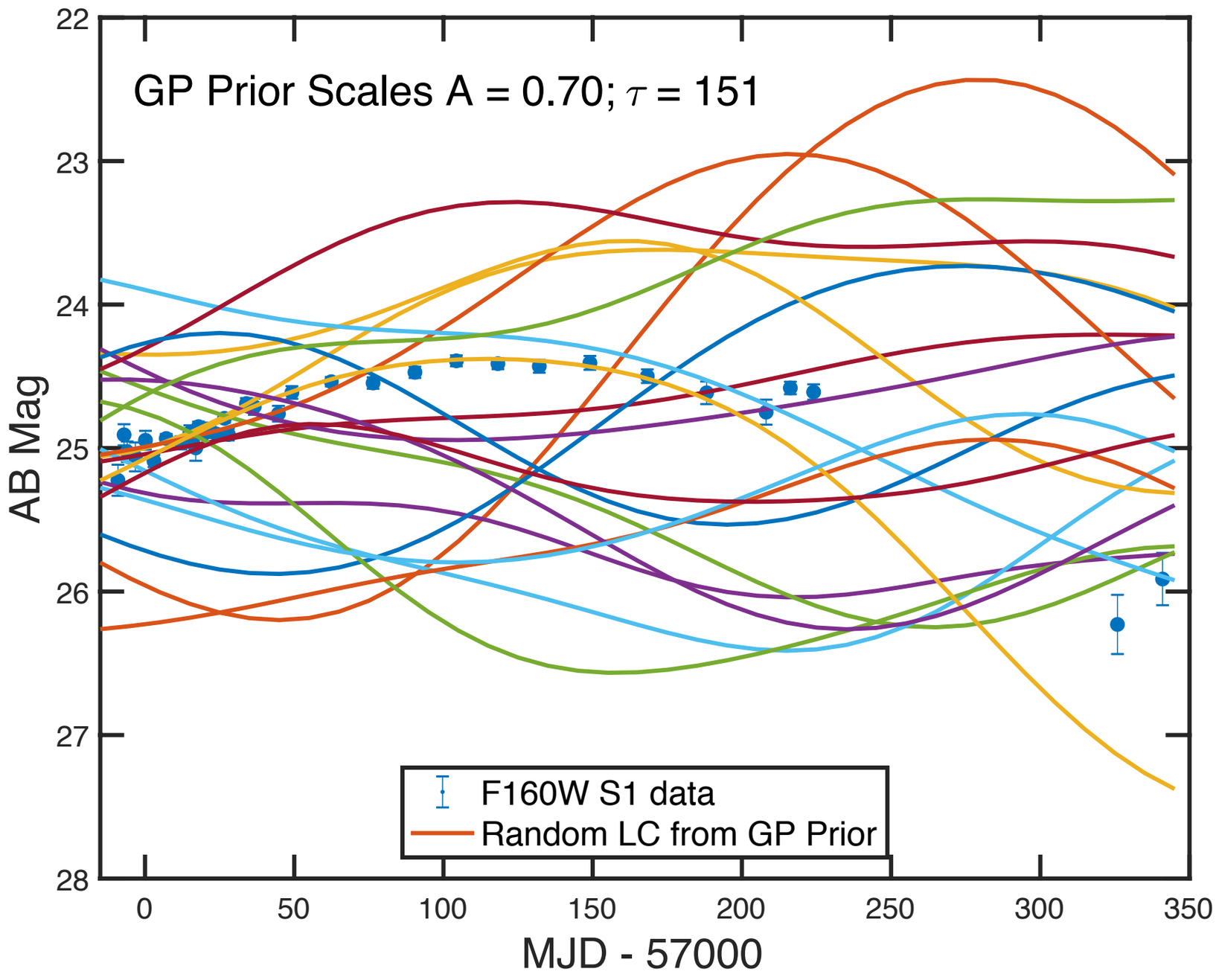
Drawing random functions from GP prior

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

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

Long Characteristic Timescale

$\tau = 151$



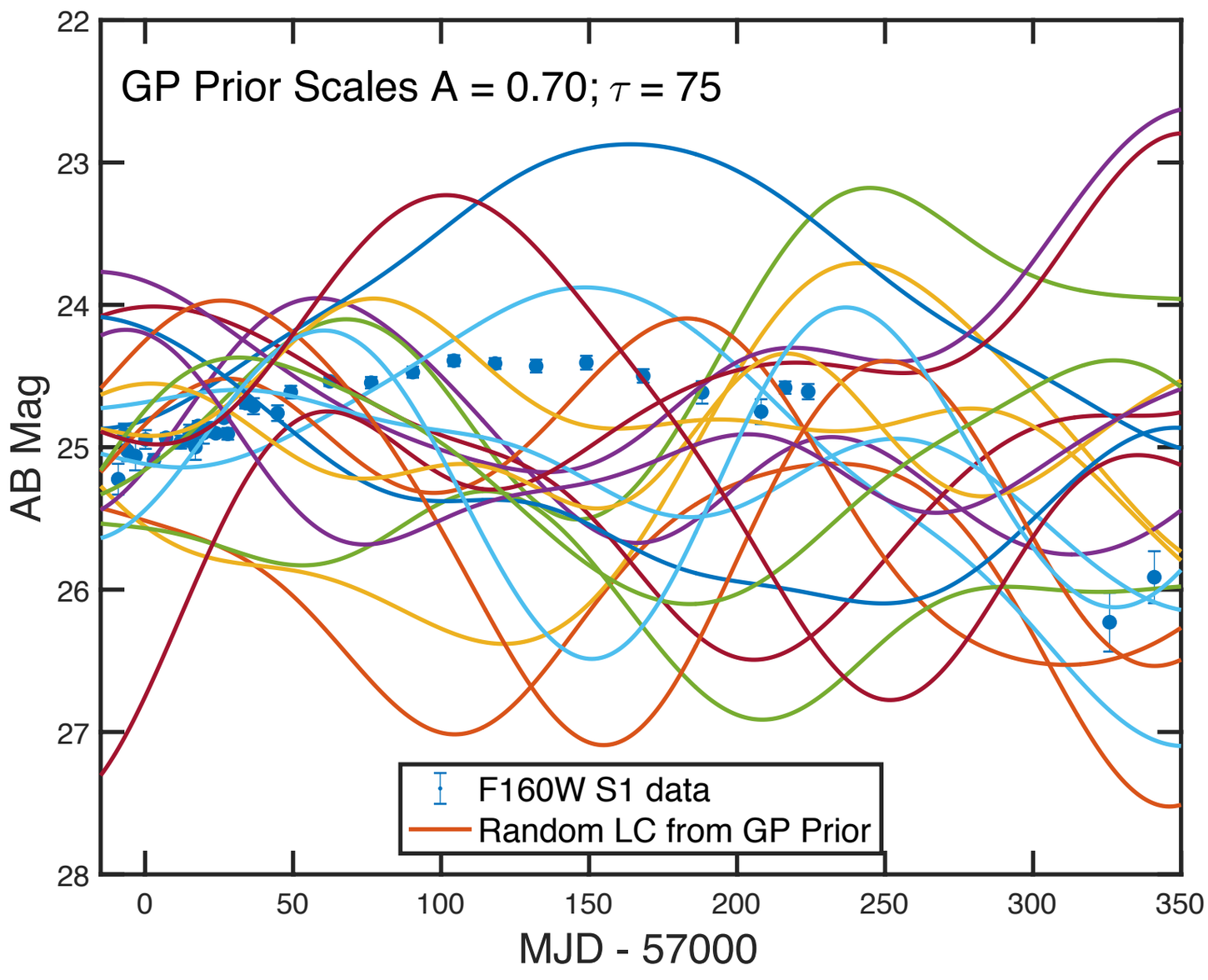
Drawing random functions from GP prior

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

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

Shorter Characteristic Timescale

$\tau = 75$



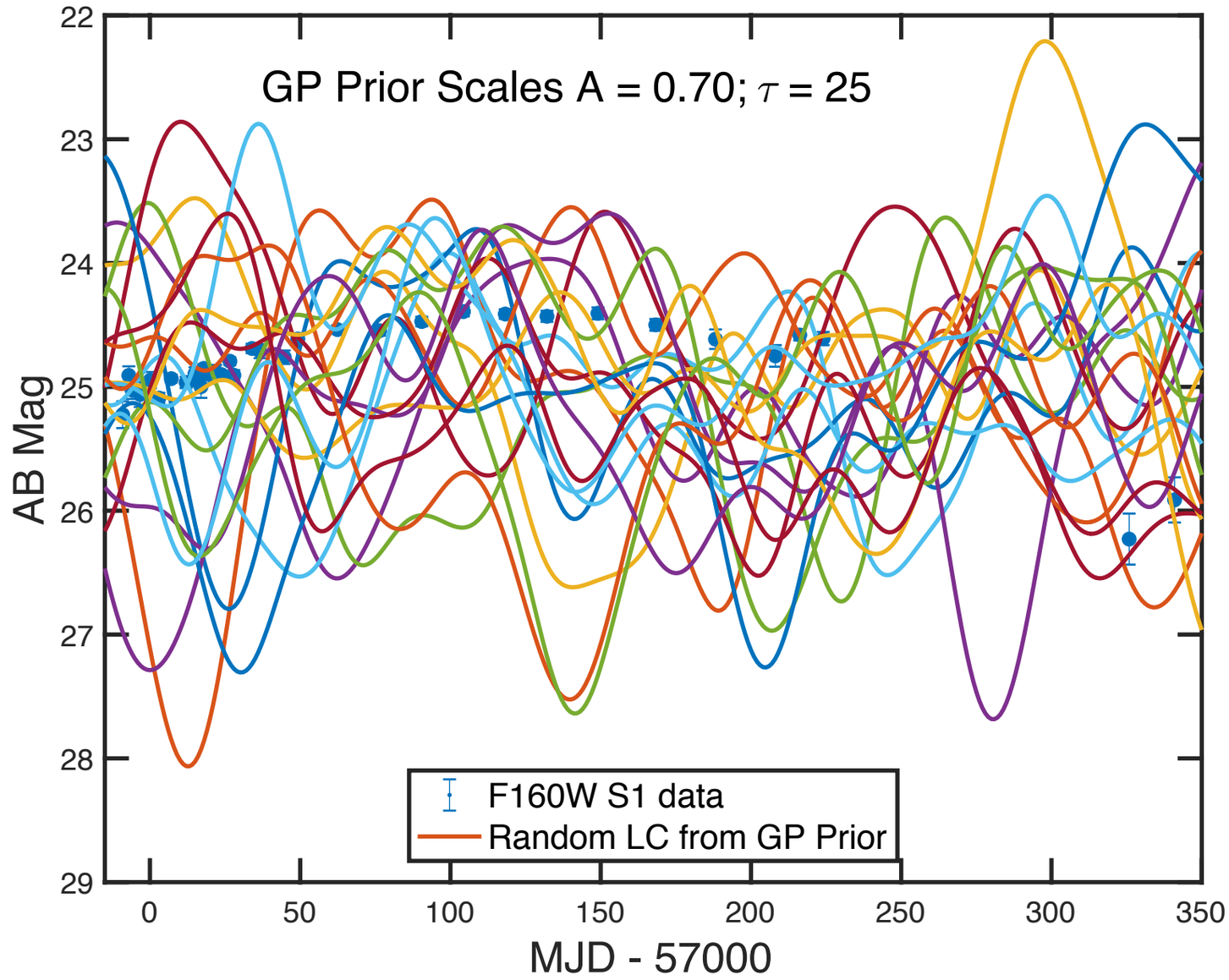
Drawing random functions from GP prior

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

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

Even Shorter Characteristic Timescale

$\tau = 25$



Fitting a GP to data

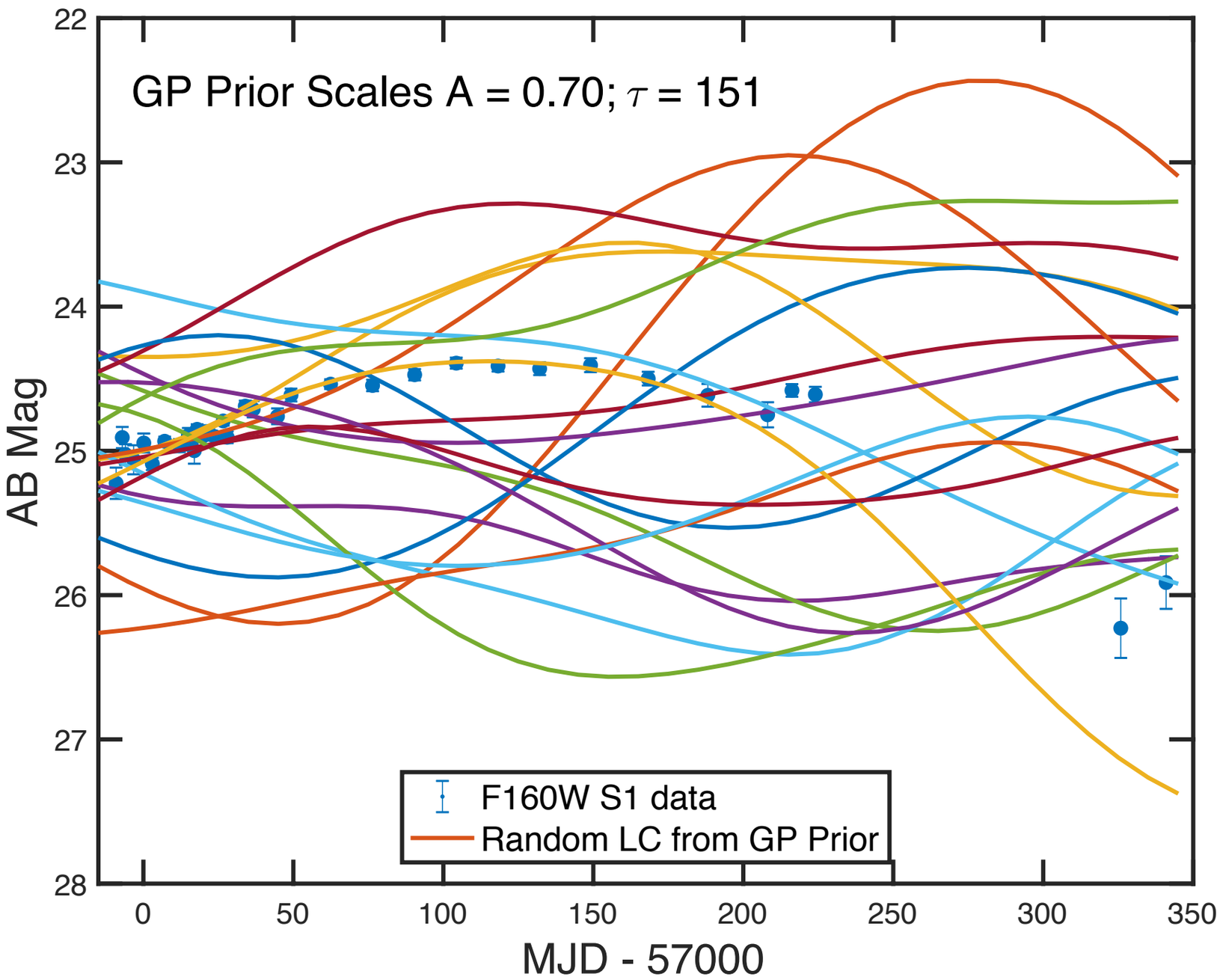
1. If we knew the characteristic scales of the kernel (A, $\boldsymbol{\tau}^2$), then how do we fit the data at observed times to find the curve for unobserved times? (computing the posterior)

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

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

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

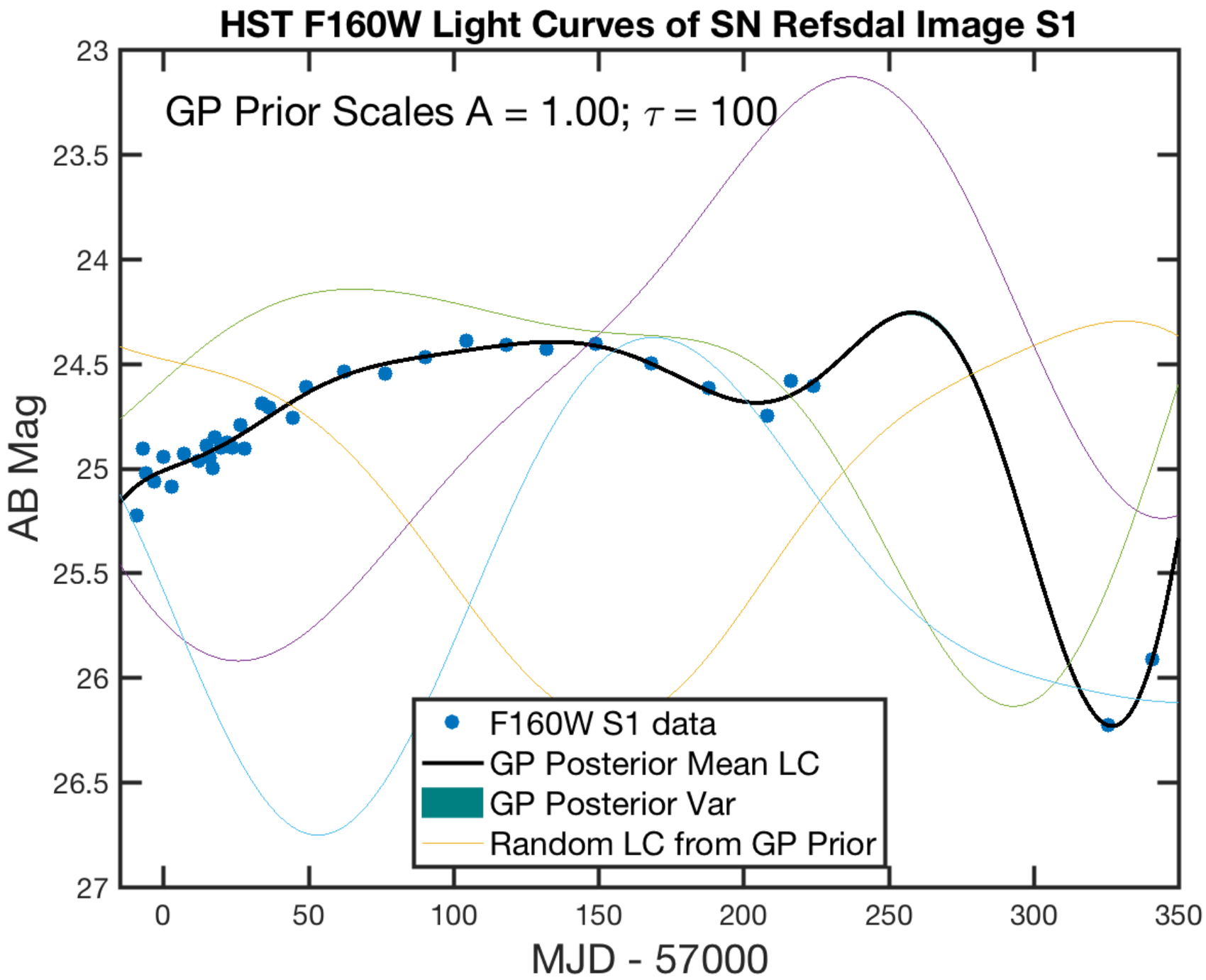
Posterior Mean:

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

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:

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