Data analysis and fitting 1

Pablo M. Reyes
SRI, Menlo Park, CA

Mainly based on slides from:
Ashton S. Reimer

ISR Summer School
July, 2025

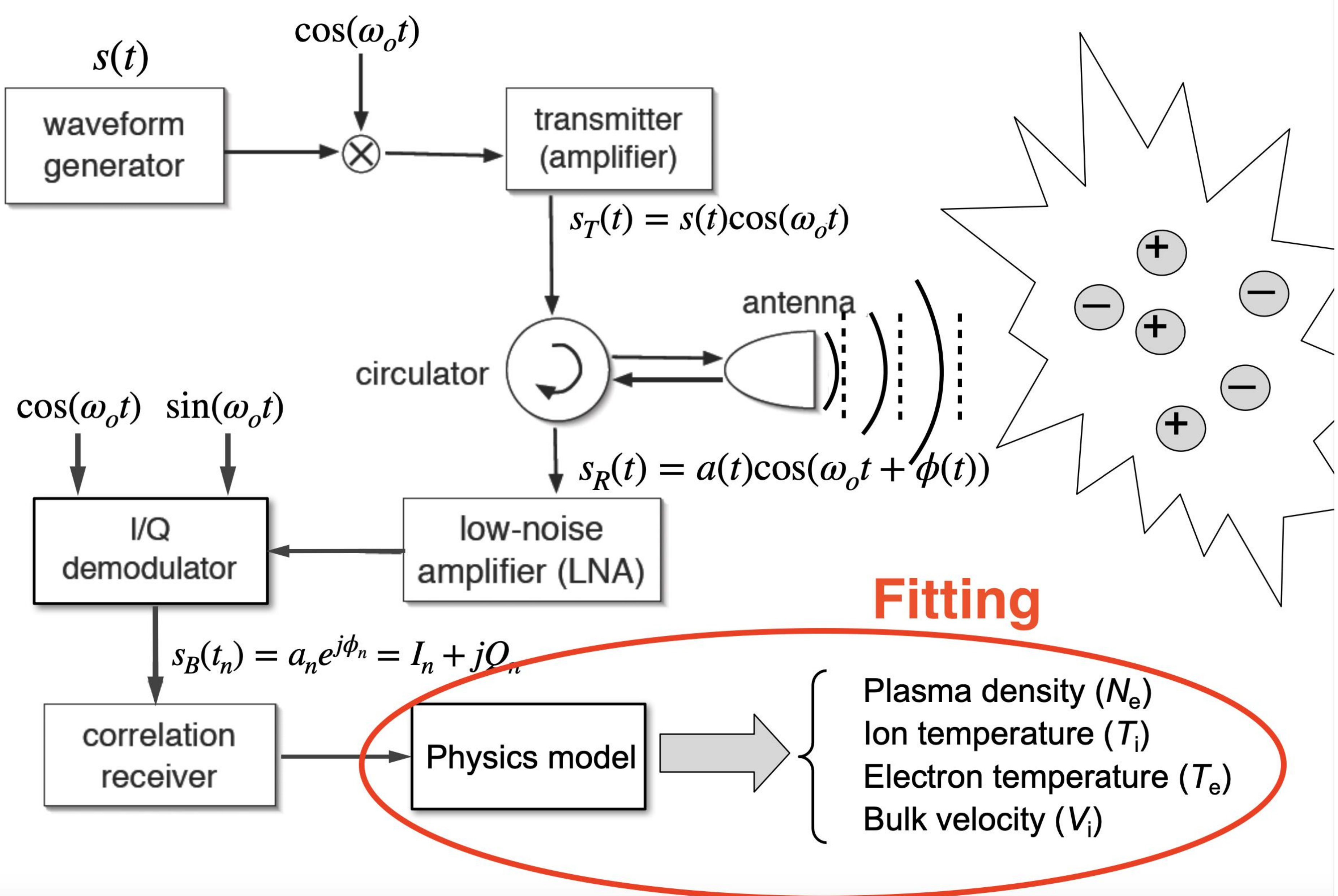
Topics to Cover

- Data Modeling
  - Forward Models
  - Inverse Problems
  - Least-Squares Technique
- Errors and Goodness of Fit:
  - Confidence Intervals and Reduced Chi-Squared
- Calibration:
  - Calibrating measurements to remove hardware bias and system unknowns.

Data analysis and fitting

- Questions:
  - What does “fitted data” mean?
  - What are the key concepts and techniques we use to “fit” data?
  - How do we go from radar signals to Ne, Te, Ti, Vlos?
  - How do I work with and interpret IS Radar data products?

Components of a Pulsed Doppler Radar



Important Concepts

- **Measurements**: directly recorded by an instrument, e.g.:
  - mercury increasing in volume within a thermometer
  - voltage samples from an analog to digital converter
- **Observations**: estimated from measurements using physics:
  - thermal expansion
  - IS Radar theory
- A **forward model**, f, contains the physics to predict measurements, $\mathbf{y}$, given observations, $\mathbf{p}$:
  - $\mathbf{y} = f(\mathbf{p})$
  - $\mathbf{p} = p_1, p_2, ..., p_N$ $\quad\quad$ $\mathbf{y} = y_1, y_2, ..., y_M$
- The **inverse problem**: With f and $\mathbf{y}$, how do we obtain $\mathbf{p}$?
  - It is rarely as simple as: $f^{-1}(\mathbf{y}) = \mathbf{p}$
  - Measurements are noisy: $\mathbf{y} = f(\mathbf{p}) + \mathbf{e}$
  - This usually requires making assumptions!

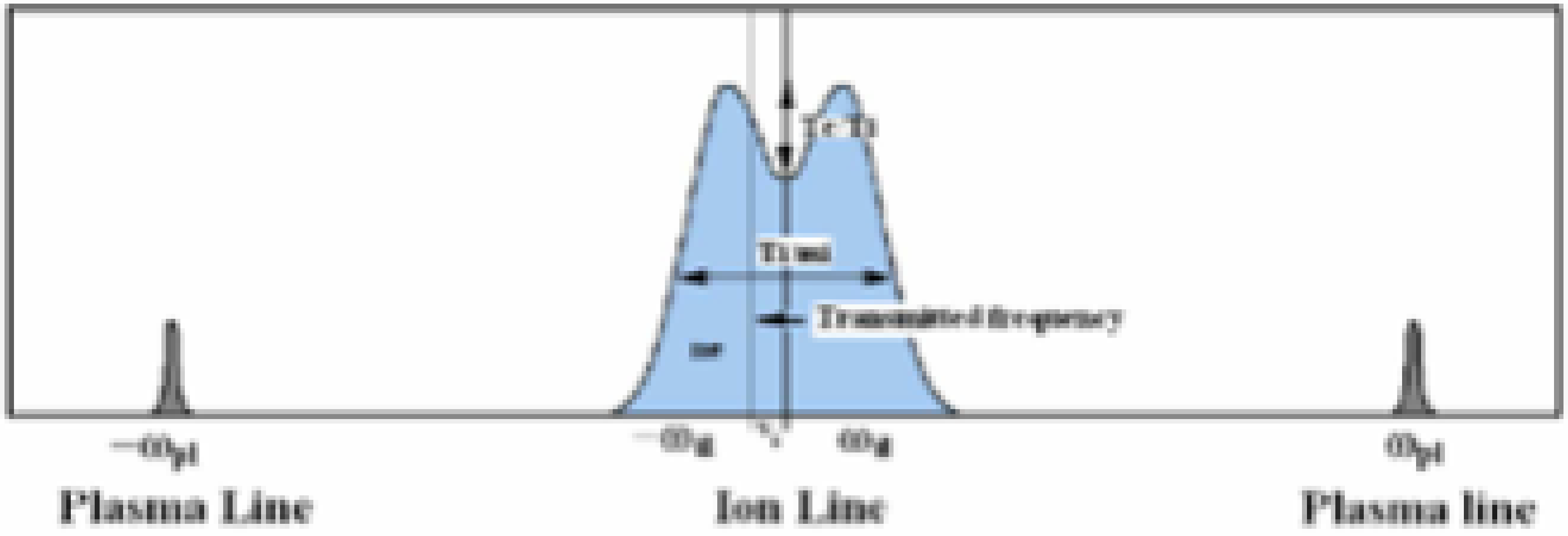
IS Radar Forward Model: Theory

IS Radar theory predicts the statistical properties of the Power Spectrum of the scattered signal:

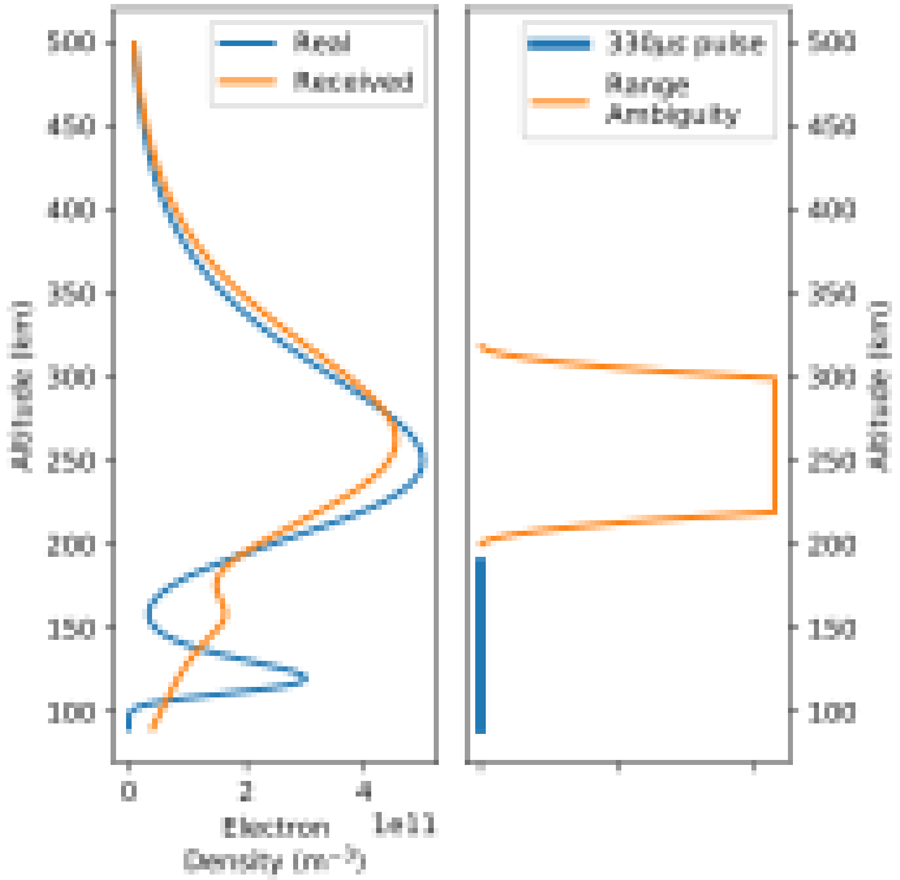
$$\langle |n_e(\mathrm{k}, \omega)|^2 \rangle \xleftrightarrow{\mathcal{F}} \langle E_s(t) E_s^*(t-\tau) \rangle$$

as a function of plasma parameters: N_e, T_e, T_i, v_{LOS}.

Forward Model? $f(N_e, T_e, T_i, v_{LOS}) = \langle E_s(t) E_s^*(t-\tau) \rangle$



IS Radar Forward Model: Ambiguity



Ambiguity:
How we measure causes "blurring" in space & time

- $W_{t\tau}$: "range-lag ambiguity function"

Forward model:
$f(N_e, T_e, T_i, v_{LOS}, W_{t\tau}) = \langle E_s(t) E_s^*(t-\tau) \rangle$

Full Forward Model Including External Noise

$$\langle V^* (t_{s1}) V (t_{s2}) \rangle = \int d\tau dr \, A (r, \tau) W_{t_{s1}, t_{s2}} (\tau, r) + R_N (t_{s2} - t_{s1})$$

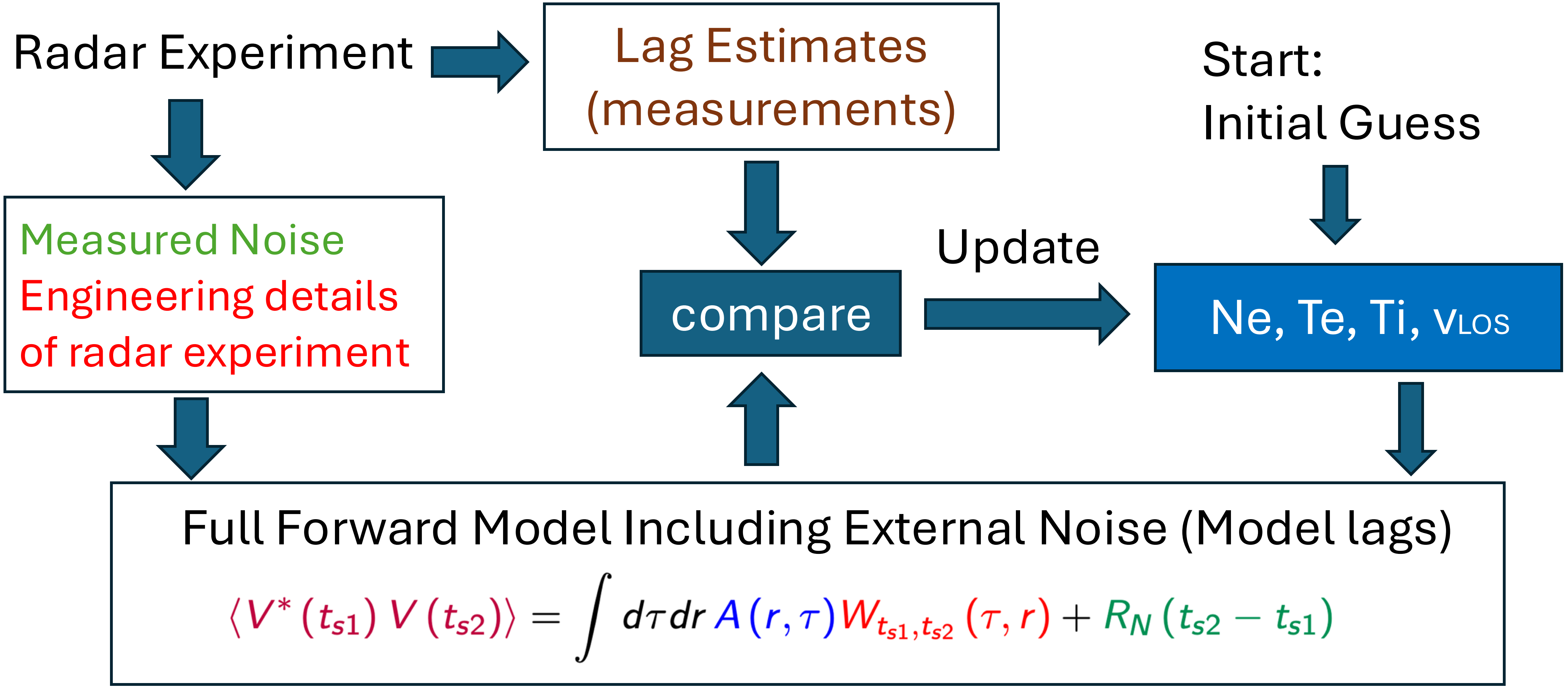
Measured elements of the LPA are determined by

- Physics of the ionosphere (unknown)
- Engineering details of radar experiment (known)
- ACF of external noise (estimate from long ranges)

Goal of fitting is to go backwards from measurements to parameters determining the physics of the ionosphere.

Fitting as an Iterative process

Goal: Obtain ionospheric parameters from radar measurements.

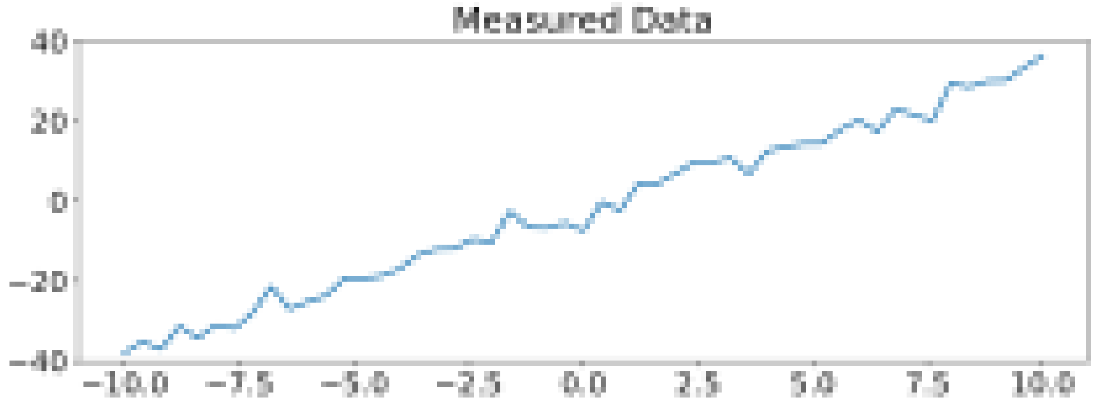


Measurement Uncertainty

Inverse problem:

- What if we have a forward model, f, and measurements, y? How do we get the observables, p?

We can try $f^{-1}(\mathrm{y}) = \mathrm{p}$, but what if measurements are noisy:



Perhaps a better forward model is $\mathrm{y} = f(\mathrm{p}, \mathrm{e}) = g(\mathrm{p}) + \mathrm{e}$, but how do we invert this?

Least-Squares Estimation

For M data points, y_m, with independent measurement errors, σ_m, compute the "chi-square"; an error weighted difference between the data and the model, f_m:

$$\chi^2(\mathrm{p}) = \sum_{m=1}^{M} \frac{[y_m - f_m(\mathrm{p})]^2}{\sigma_m^2}$$

the model parameters that provide the "best fit" of the model to the data, $\hat{\mathrm{p}}_{LS}$, are those that minimizes $\chi^2(\mathrm{p})$: $\underset{\mathrm{p}}{\operatorname{argmin}} \left\{ \sum_{m=1}^{M} \frac{[y_m - f_m(\mathrm{p})]^2}{\sigma_m^2} \right\}$

In general, measurements y_m may not be independent. The generalized least-squares estimate is:

$$\chi^2(\mathrm{p}) = [\mathrm{y} - f(\mathrm{p})]^T \, \Sigma_e{}^{-1} \, [\mathrm{y} - f(\mathrm{p})]$$

where Σ_e is the covariance matrix of measurements y.

Least-Squares Estimation

Important terminology and concepts:

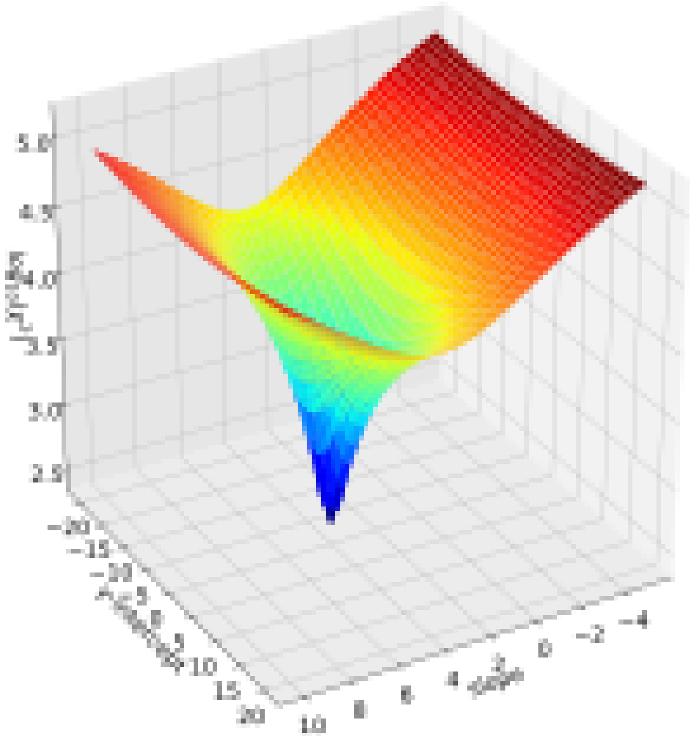
- $\frac{y_m - f_m(\mathrm{p})}{\sigma_m}$ are known as the "normalized errors". The numerator is called the "residual".
- assumed that each "normalized error" is Gaussian distributed with zero mean and unit variance: $\mathcal{N}(0,1)$
- the chi-squared, χ^2, is the sum of the square of $\mathcal{N}(0,1)$ random variables, so by definition, χ^2 is a chi-squared distributed random variable, with $M - N$ degrees of freedom.

Example: Fitting a Linear Model to Data

Given a Model:

$$y = mx + b$$

Calculate $\chi^2(m, b)$:

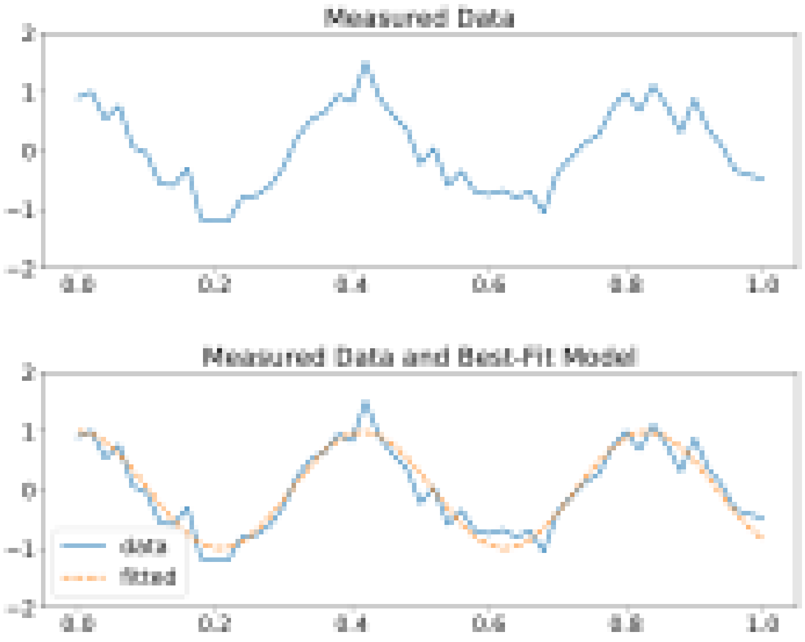


Notebook: data_analysis_and_fitting/02_modeling_and_data_fitting/Fitting Linear Data Example.ipynb

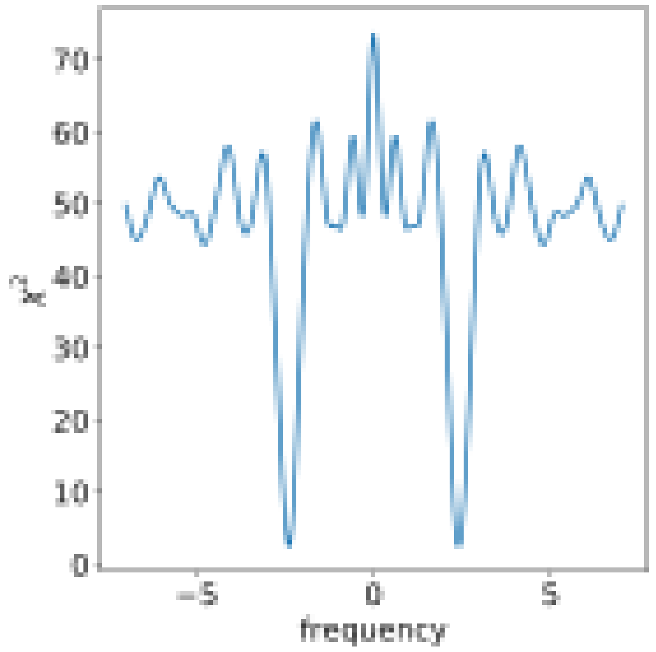
Example: Fitting a Non-Linear Model to Data

Given a Model:

$$y = \cos(2\pi f t)$$



Calculate $\chi^2(f)$:



Notebook: data_analysis_and_fitting/02_modeling_and_data_fitting/Fitting Linear Data Example.ipynb

Least-Squares Estimation

For some simple models (linear, quadratic, etc):

- analytic solutions exist for the minimum χ^2

In general, non-linear least squares algorithms are required:

- Levenberg-Marquardt (LM) algorithm is most commonly used
- LM requires a good initial guess
- Standard LM packages:
  - FORTRAN: MINPACK lmdif.f and lmder.f
  - Python: scipy.optimize.leastsq (wrapper around lmdif and lmder)
  - Matlab: Optimization Toolbox lsqnonlin
  - IDL: LMFIT

Next Topics

- Errors
- Goodness of Fit

Chi-Squared

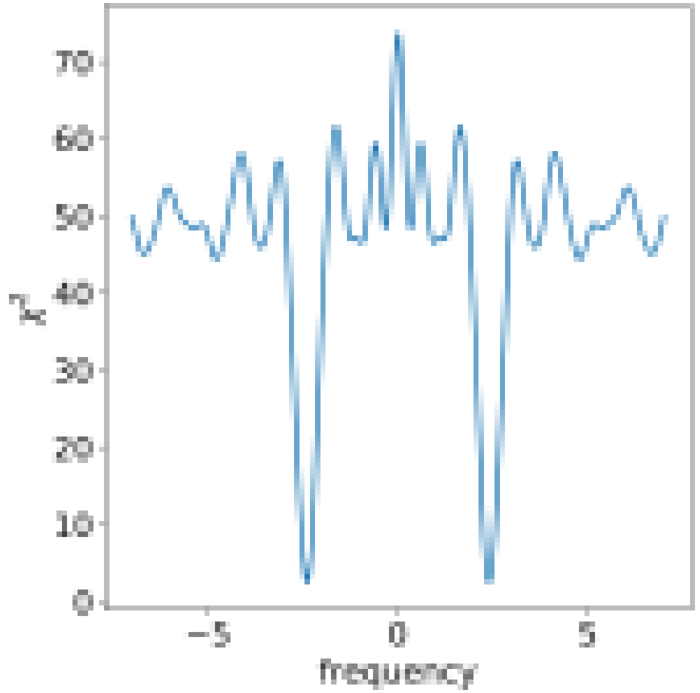
We can use least-squares to solve inverse problems:

$$\chi^2(\mathrm{p}) = [\mathrm{y} - f(\mathrm{p})]^T \, \Sigma_e^{-1} \, [\mathrm{y} - f(\mathrm{p})]$$

where $\hat{\mathrm{p}}_{LS}$ are the "best-fit" model parameters, those that minimizes $\chi^2(\mathrm{p})$

Great! But:

- What are the errors in the fitted parameters $\hat{\mathrm{p}}_{LS}$?
- Is the fit meaningful? Does the model accurately reproduce the measurements?



Error Propagation (e.g. Linear Least-Squares)

For a linear forward model:

$$\mathrm{y} = f(\mathrm{p}) + \mathrm{e} \qquad f(\mathrm{p}) = H\mathrm{p}$$

The Least-Squares solution is:

$$\hat{\mathrm{p}}_{LS} = \left[H^T \Sigma_e^{-1} H\right]^{-1} H^T \Sigma_e^{-1} \mathrm{y}$$

Given that jointly Gaussian random variables have the following property:

$$\mathrm{Y} = A\mathrm{X} \quad \Rightarrow \quad \Sigma_{\mathrm{Y}} = A \Sigma_{\mathrm{X}} A^T$$

it can be shown that:

$$\Sigma_{\hat{\mathrm{p}}_{LS}} = \left[H^T \Sigma_e^{-1} H\right]^{-1}$$

Error Propagation (e.g. Nonlinear Least Squares)

For a non-linear forward model, guess a p_i, linearize, and step towards minimum:

$$y = f(p) + e \qquad f(p_i + \Delta p) \approx f(p_i) + J_i \Delta p \qquad J_i = \frac{\partial f}{\partial p_i}$$

J is known as the Jacobian:

$$J = \begin{pmatrix} \frac{\partial f_0}{\partial p_0} & \frac{\partial f_0}{\partial p_1} & \cdots & \frac{\partial f_0}{\partial p_{N-1}} \\ \frac{\partial f_1}{\partial p_0} & \frac{\partial f_1}{\partial p_1} & \cdots & \frac{\partial f_1}{\partial p_{N-1}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f_{M-1}}{\partial p_0} & \frac{\partial f_{M-1}}{\partial p_1} & \cdots & \frac{\partial f_{M-1}}{\partial p_{N-1}} \end{pmatrix}$$

J is $M \times N$ (tall and skinny)

Non-linear fitting process:

- iterate until $p_{i+1} = \hat{p}_{LS}$: that which minimizes χ^2
- The covariance of $\hat{p}_{LS}$ is:

$$\Sigma_{\hat{p}_{LS}} = \left[J^T \Sigma_e^{-1} J \right]^{-1}$$

Note the similarity to the linear case!

Error Propagation

The covariance of the fitted parameters is the covariance of the input data propagated through the least-squares operation:

$$\Sigma_{\hat{\beta}_{LS}} = \left[J^T \Sigma_e^{-1} J \right]^{-1}$$

"Error bars" for fitted parameters:

- Assumption: measurement errors are **Gaussian** distributed with covariance Σ_e, denoted $\mathcal{N}(0, \Sigma_e)$
- The "errors" in the fitted parameters are related to confidence intervals
- Confidence intervals are constructed from $\Sigma_{\hat{\beta}_{LS}}$
- $\Sigma_{\hat{\beta}_{LS}}$ may look reasonable, even if the fit is meaningless

Constructing Confidence Intervals: From Fitted Covariance

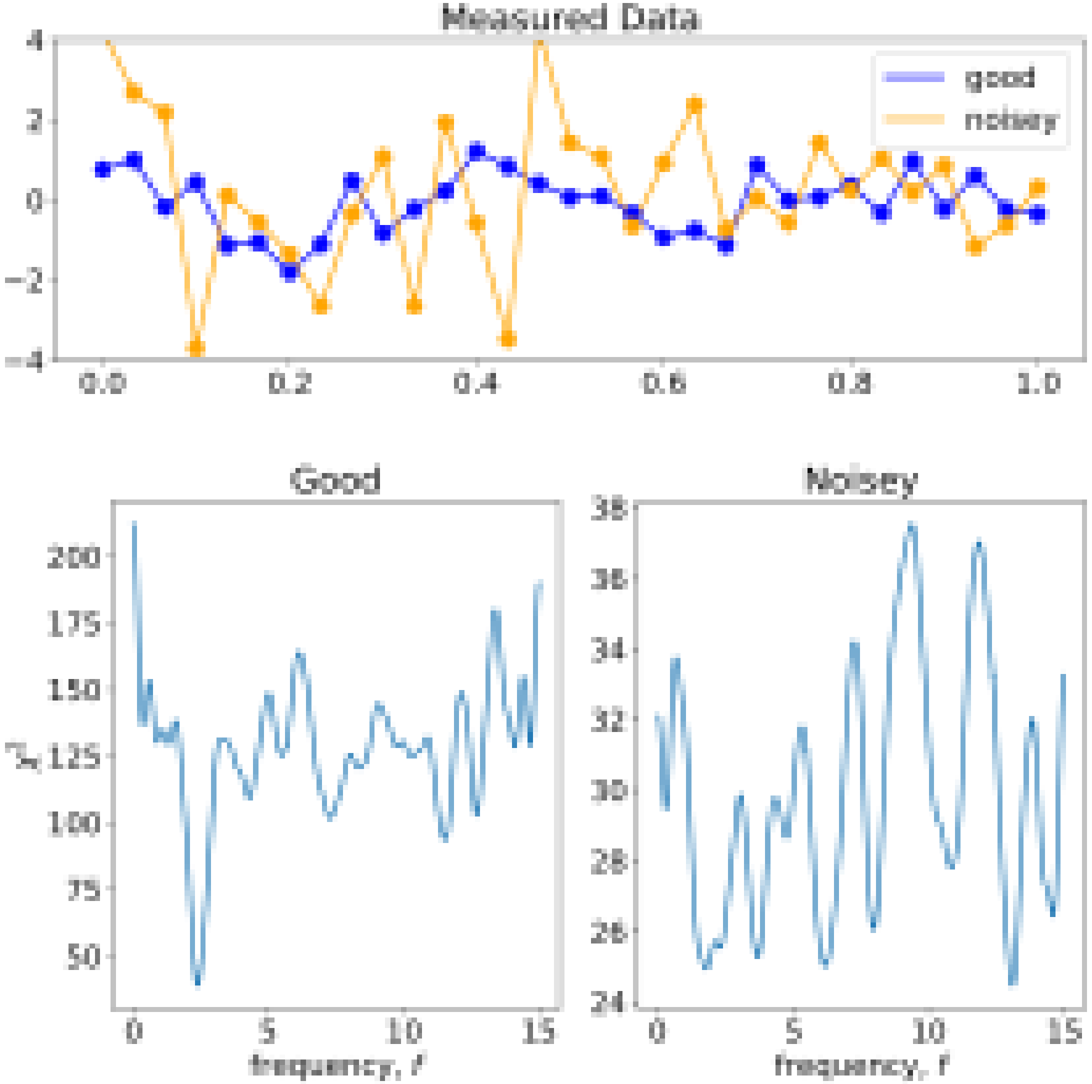
Error bars, δp_m, for a fitted parameter can be constructed from the covariance $\Sigma_{\hat{p}_{LS}}$ and a $\Delta\chi^2$:

$$\delta p_m = \pm\sqrt{\Delta\chi^2}\sqrt{\Sigma_{mm}}$$

The value of $\Delta\chi^2$ selects the "significance level":

- $\Delta\chi^2$ is found in lookup tables calculated from the CDF of the χ^2 distribution
- Single parameter fit, $N = 1$:
  - a 68% significance: $\Delta\chi^2 = 1$
  - a 95.4% significance: $\Delta\chi^2 = 4$
- Two parameter fit, $N = 2$:
  - a 68% significance: $\Delta\chi^2 = 2.3$

Challenges With Constructing Confidence Intervals



Validity of Confidence Intervals

Only quantitatively valid when:

- measurement errors are Gaussian, and
  - the model $f(p)$ is linear in for all p, or
  - measurement errors are small enough that $f(p)$ can be accurate approximated by a linear model in the region around p

Otherwise, alternative fitting methods are required: Monte Carlo, Bayesian, etc.

Goodness of Fit

How do we know if the fit is even meaningful? The standard goodness of fit test involves computing the "reduced chi-squared":

$$\chi^2_\nu = \chi^2/(m - n + 1)$$

Then, typically:

- $\chi^2_\nu \approx 1$: a good fit
- $\chi^2_\nu << 1$: an "over fit"
- $\chi^2_\nu >> 1$: a poor fit

The χ^2_ν could also be slightly larger or smaller than 1 depending on how accurately one is able to estimate the input measurement errors.

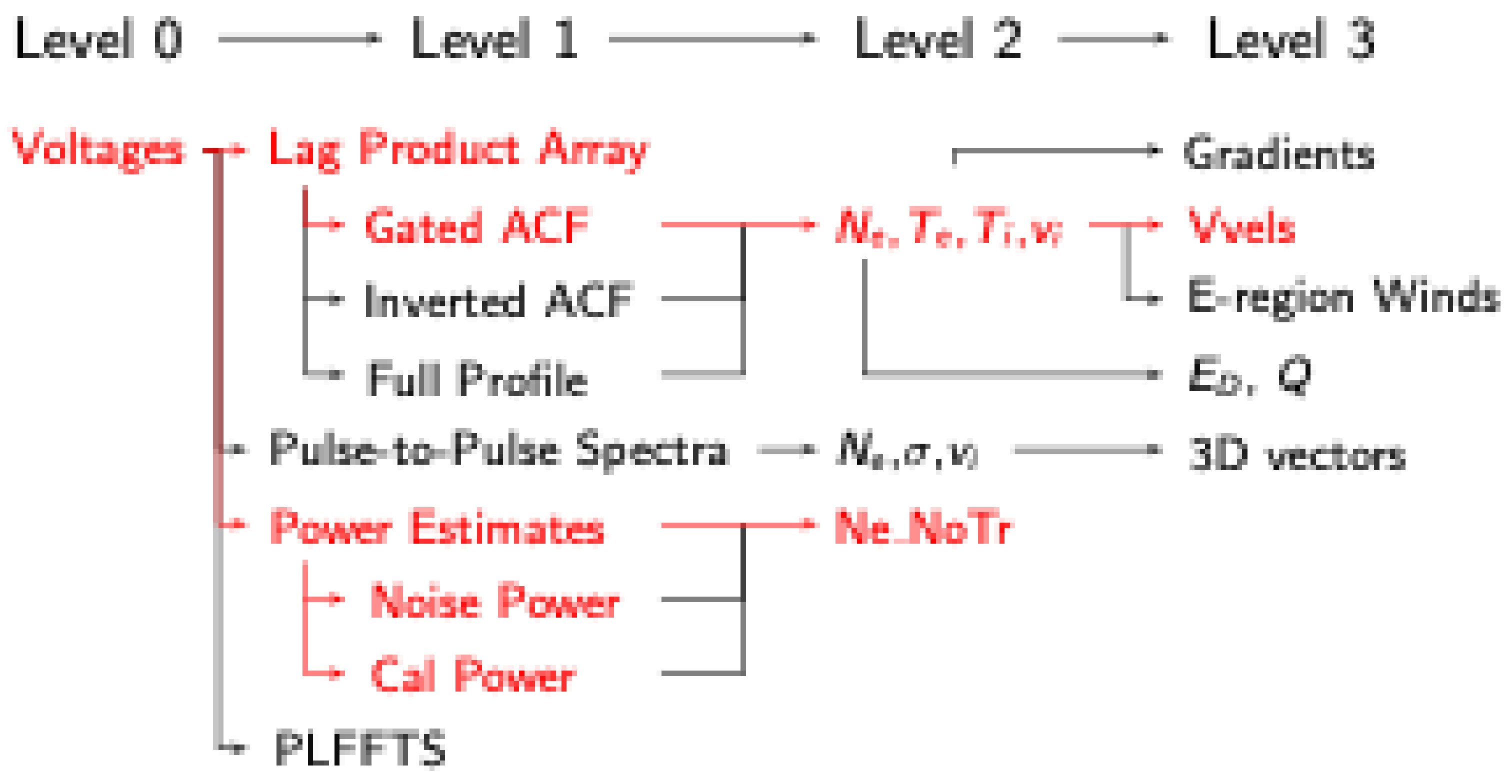
Summary

Now we can answer the question: Are the fitted parameters meaningful?

- What is the uncertainty in the fitted parameters?
  - Error bars correspond to confidence intervals (CI)
  - CIs are constructed from covariance of the fitted parameters
  - For a 68% CI, interpretation is: "If we could hypothetically make and infinite set of new measurements and fit each of those, 68% of the time the 'true' value of the parameter would lie within the CI."
- Is the fit good?
  - Compute the reduced chi-squared
  - $\chi^2_\nu \approx 1$: usually means the model accurately represents the data
- All of this error analysis depends on the assumption that measurement errors are **Gaussian** distributed with covariance Σ_e such that $(y_m - f_m)/\sigma_m$ are $\mathcal{N}(0,1)$

IS Radar Data Levels

Summary of IS Radar data products:



Calibration

- Introduction to radar data calibration
- Review methods for calibration

The System Constant

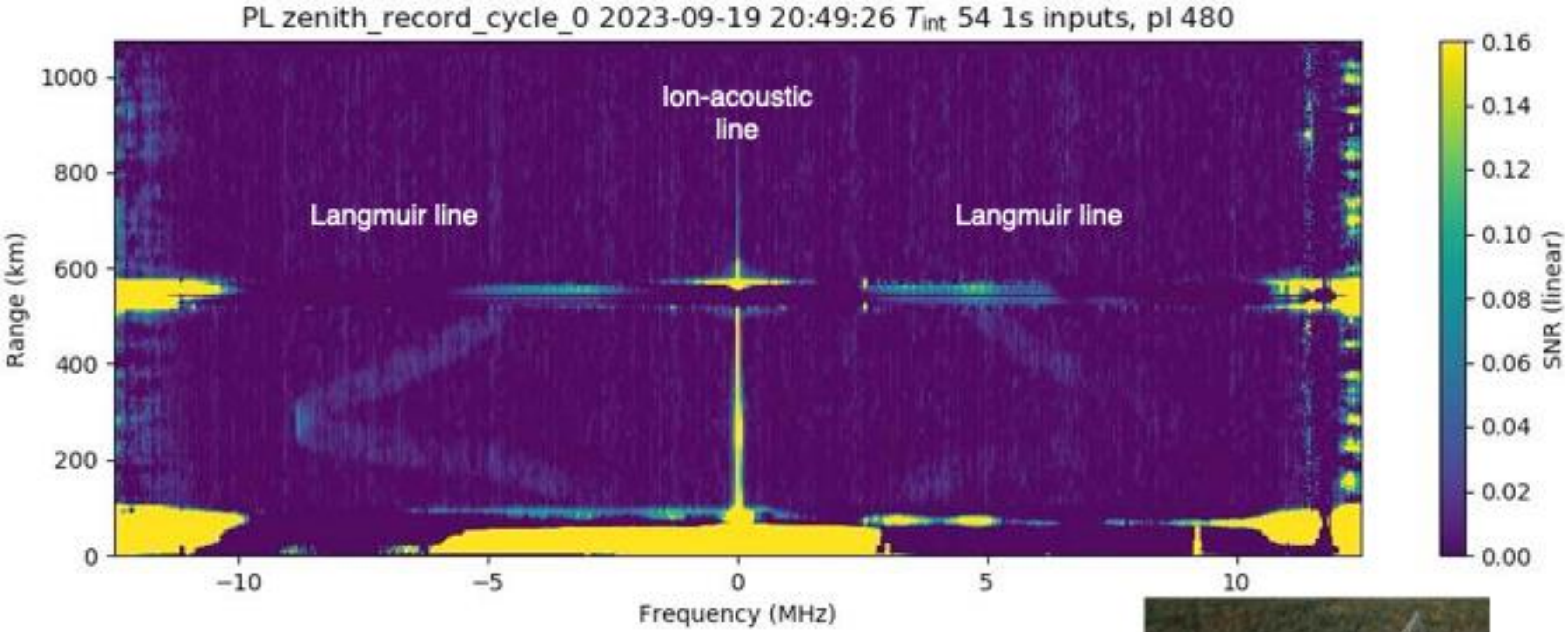
- Radar equation for IS Radar:

$$P_r = \frac{P_t \tau_p}{R^2} K_{sys} \frac{N_e}{\left(1 + k^2 \lambda_{De}^2\right)\left(1 + k^2 \lambda_{De}^2 + T_e/T_i\right)}$$

$$\lambda_{De} = \sqrt{\frac{\epsilon_0 k_B T_e}{e^2 N_e}} \quad k = \frac{4\pi}{\lambda_{\mathrm{Tx}}} \quad R = \text{Range} \quad \tau_p = \text{Pulse Length (s)}$$

- K_{sys}: the "System Constant" involves antenna gain, effective area, etc. For PFISR $K_{sys} \sim 10^{-19}\ \mathrm{m}^5\mathrm{s}^{-1}$.
- Can determine K_{sys} by comparing estimated N_e to absolute N_e measurements, e.g.:
  - Plasma line data
  - Ionosonde f_{0F2}
  - Faraday rotation (e.g. Jicamarca)
- Lags estimates need to be calibrated too!

Plasma Line Spectra: Millstone Hill (60 sec integration)

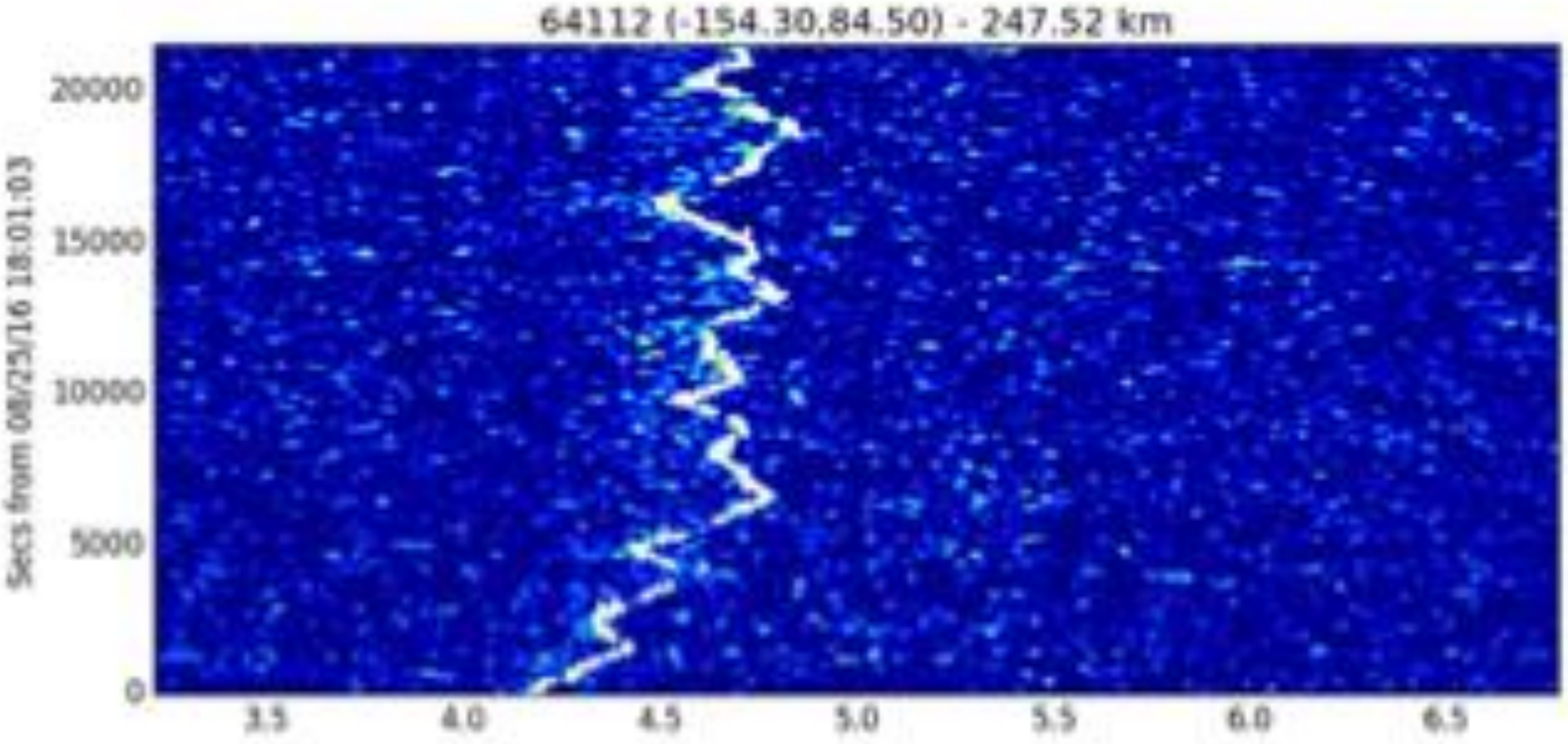


NB: waveform shape has not been deconvolved
(480 usec pulse used here)

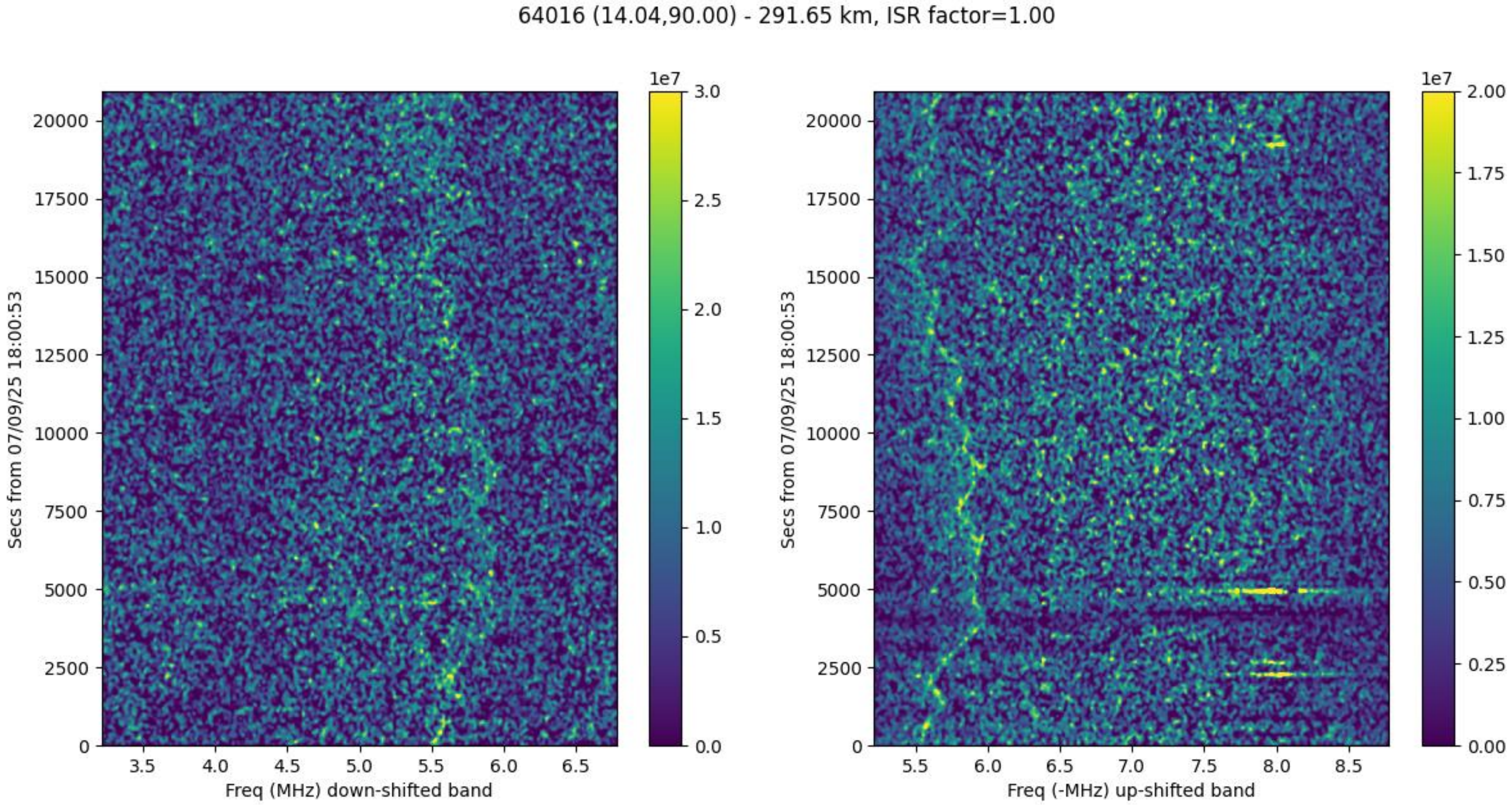
68m Zenith, 440 MHz

Enhanced Plasma Lines

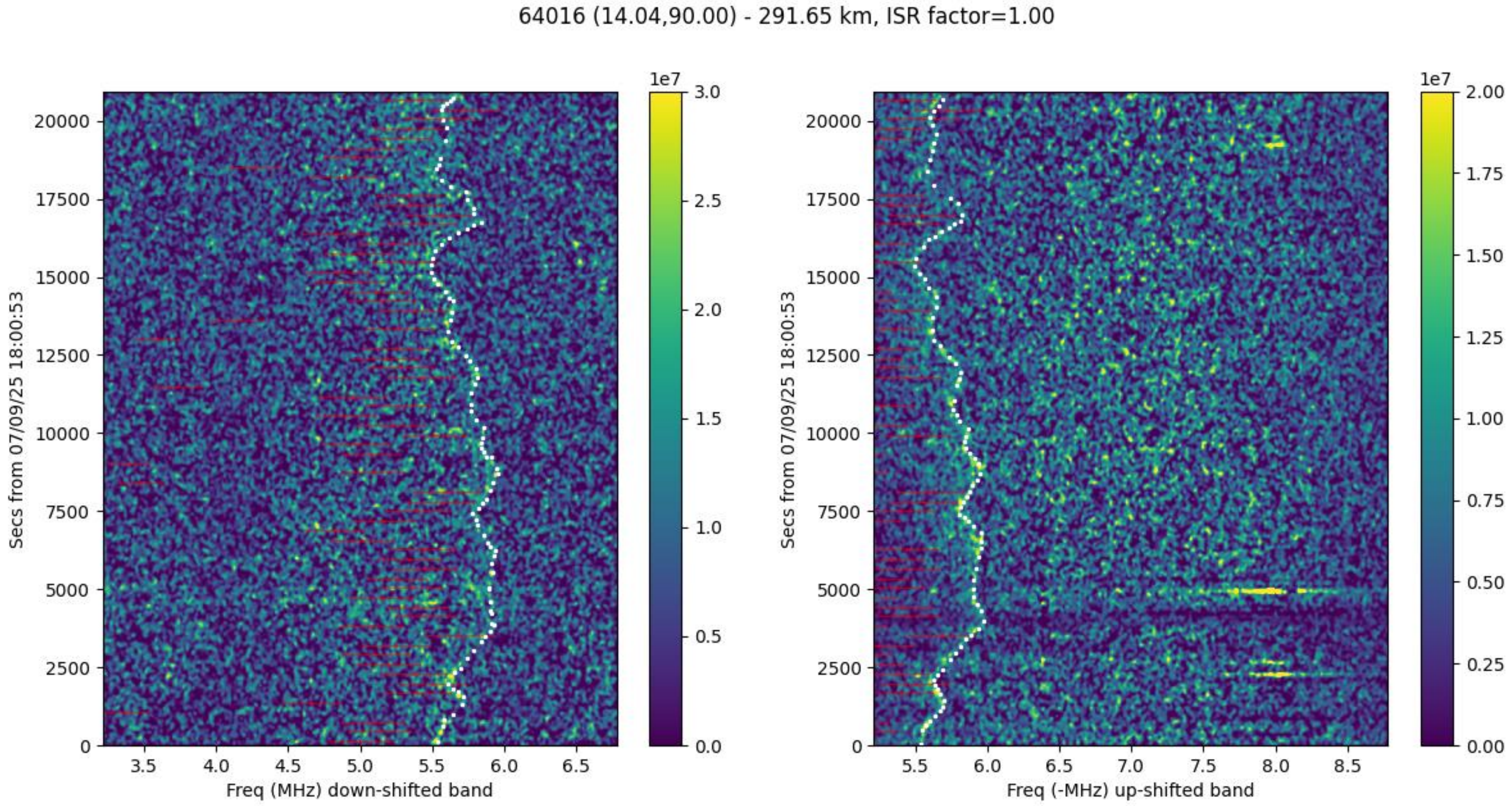
$$\omega^2 = \omega_{pe}^2 + \frac{3}{2}k^2 v_{th}^2 + \Omega_e^2 \sin^2\alpha$$



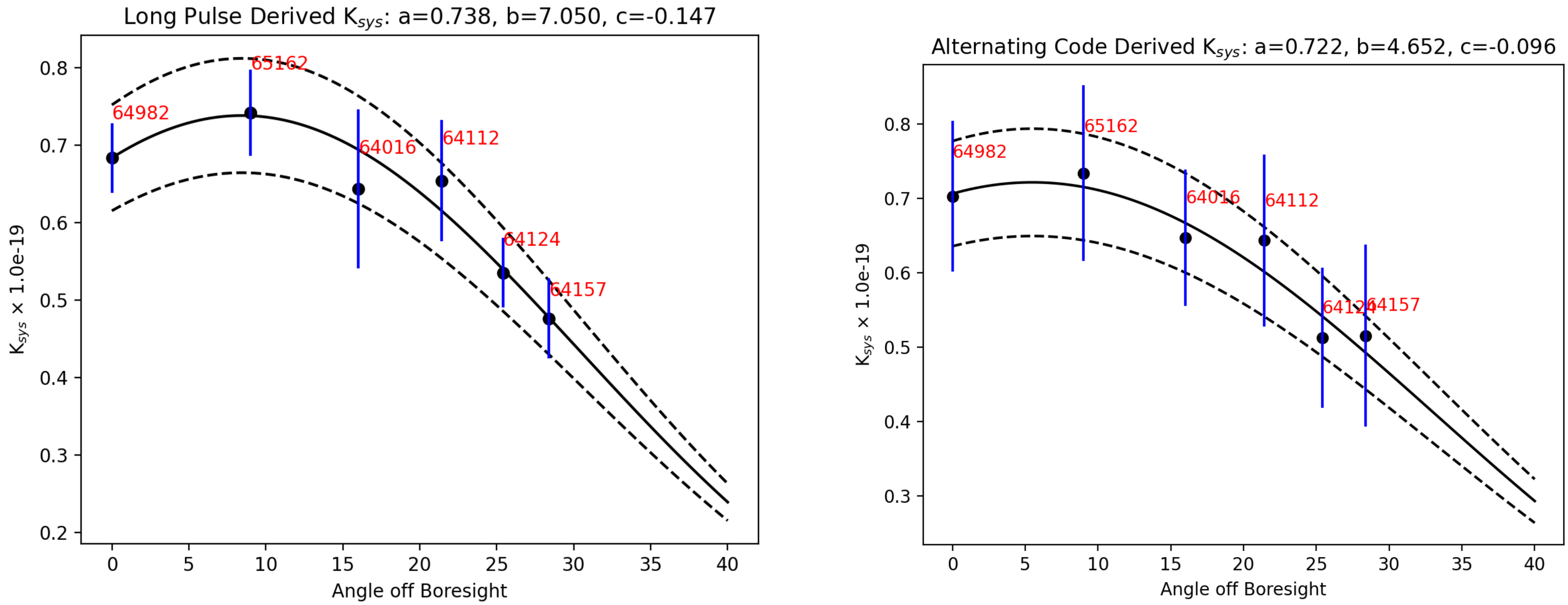
Plasma Line peak at PFISR July 2025



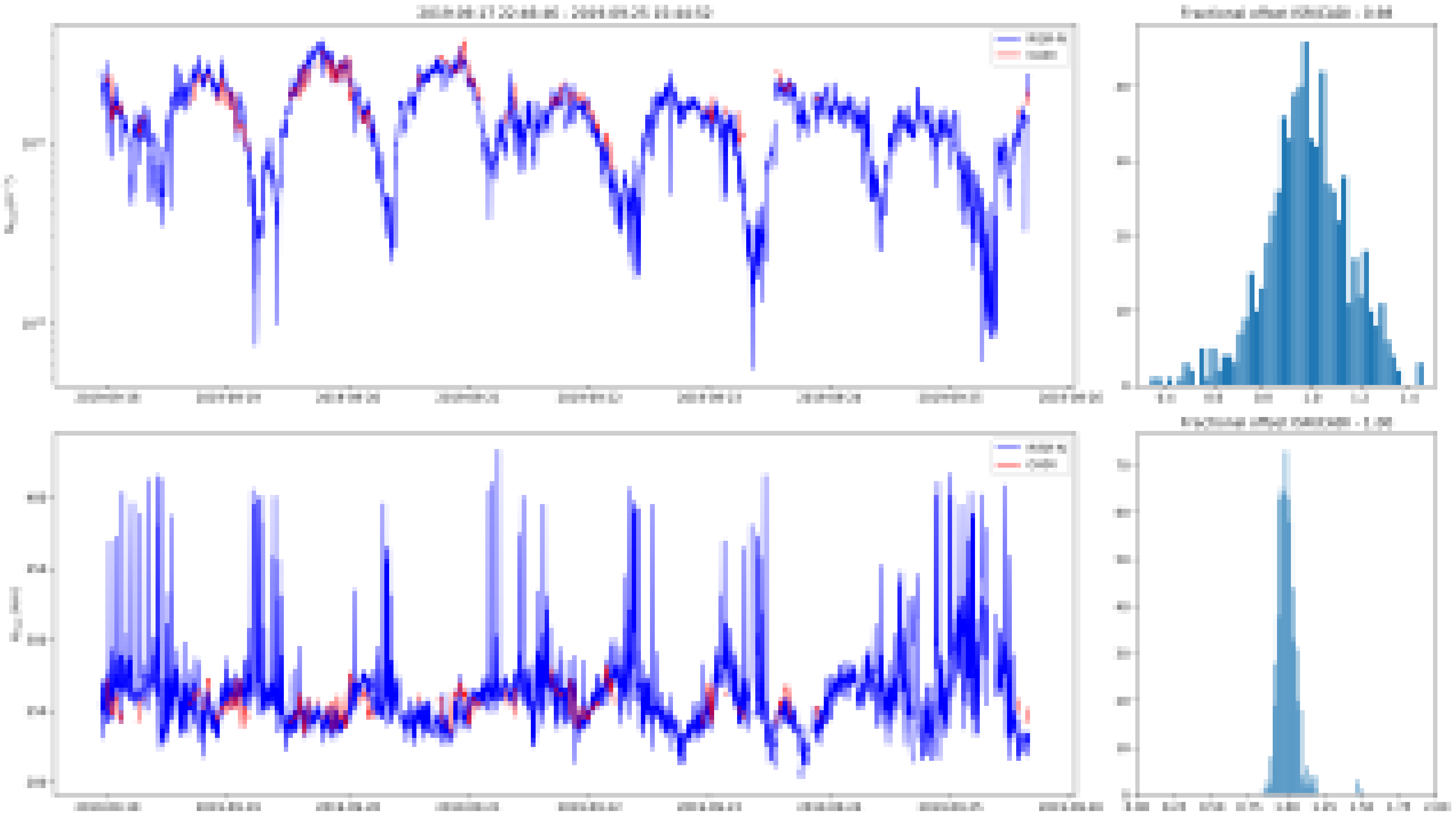
Plasma Line peak at PFISR July 2025



Plasma Line peak at PFISR July 2025



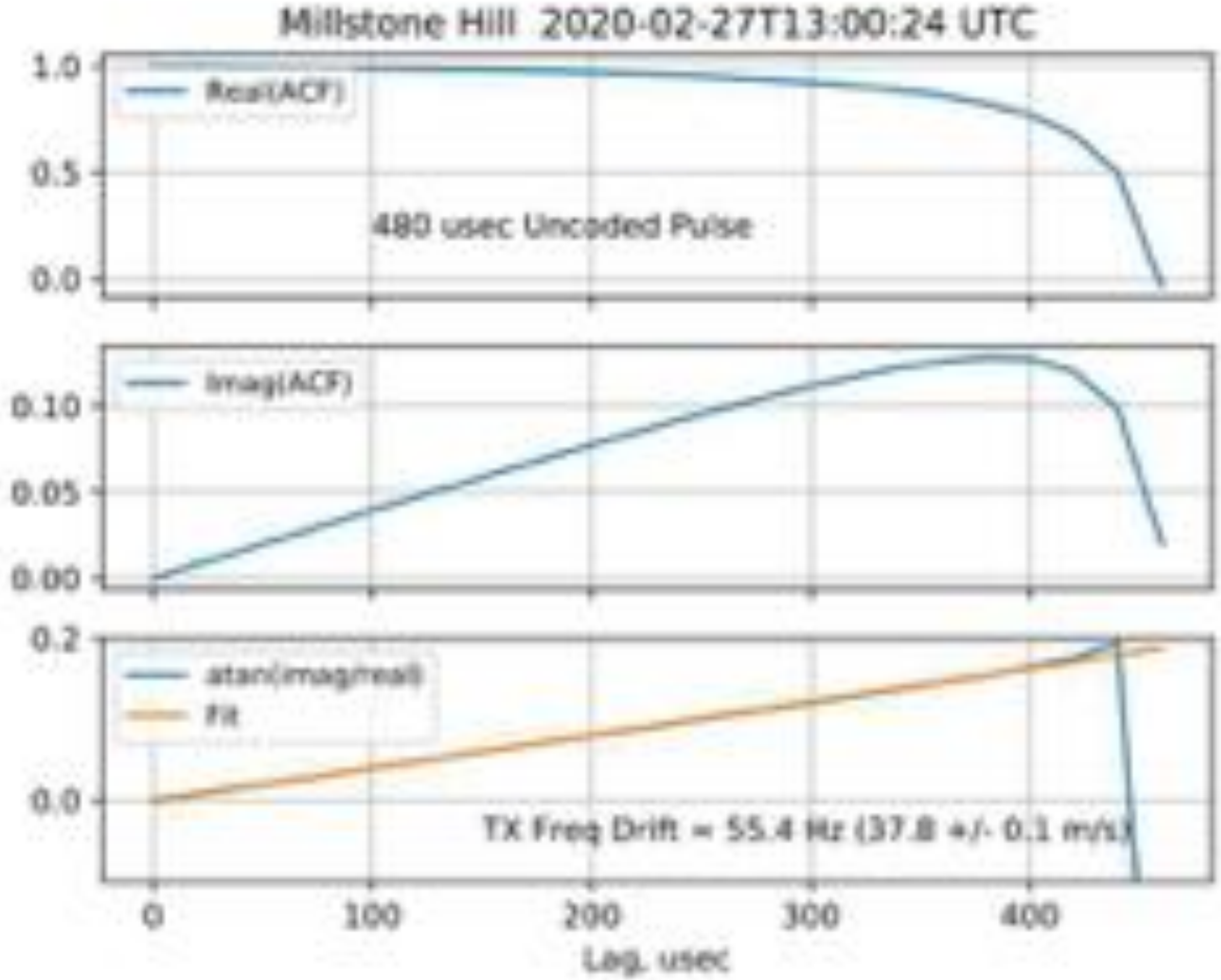
Comparisons with Ionosondes



Tx Pulse Chirp: Klystron Example

- Electrons slow down slightly as they travel through the main cavity of the klystron (TX amplifier)
- This causes a frequency offset of $\approx$0.1 ppm that will systematically bias ionospheric velocities.
- Sample Tx pulse, compute ACF, fit phase slope

Tx Pulse Chirp: Klystron Example



Summary

Radar data is obtained using real hardware and requires calibration:

- Calibration of received power into absolute electron density using:
  - Plasma line data
  - Ionosonde data
  - ...
- There may be other quirks of the hardware that need to be accounted for.

Important Distinction for Data Analysis:

- "Calibrated": Indicates that the electron density is calibrated in absolute units.
- "Uncorrected": Electron density obtained without the $(1 + T_e/T_i)^{-1}$ factor.