# (Mis)use of linear least-squares regression, and some other thoughts

Andrew Sayer, GESTAR-USRA/NASA GSFC

andrew.sayer@nasa.gov

with input from Kirk Knobelspiesse, NASA GSFC

# A first note

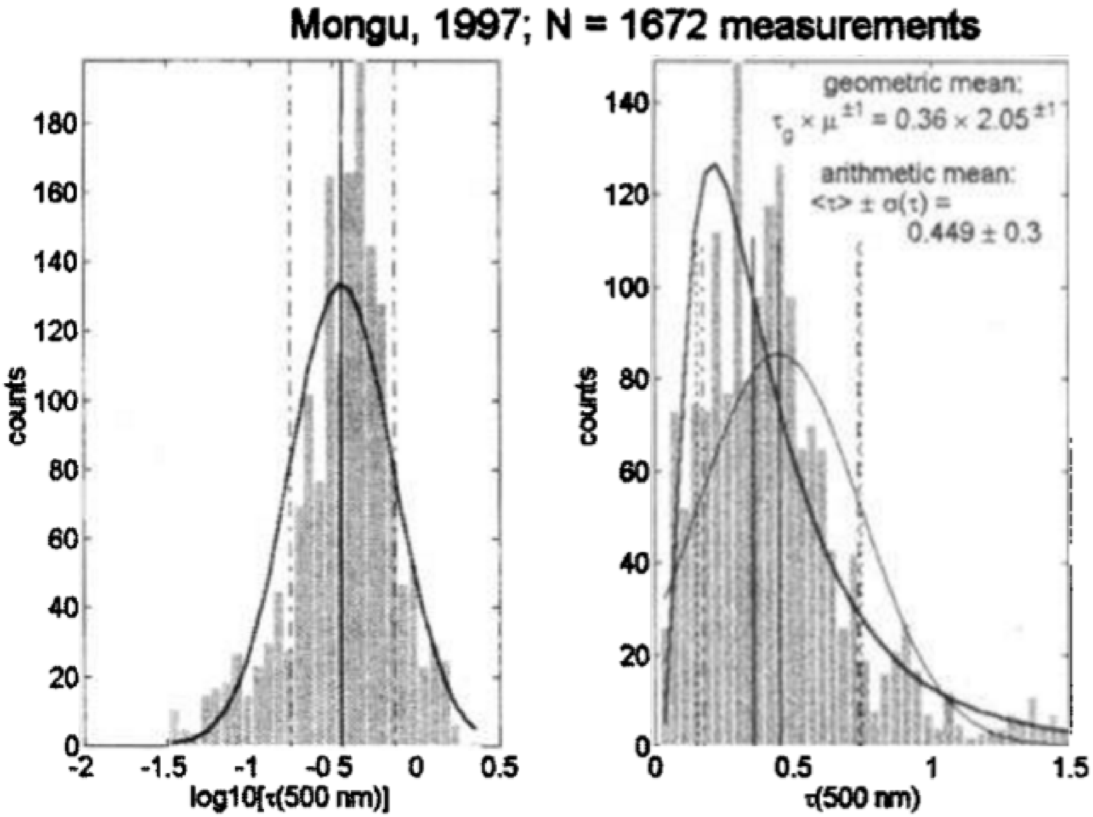
- My understanding is incomplete, but I know enough to know we’ve sometimes been doing it wrong
- The goal is **not** to name or shame
- Highlight some statistical difficulties with the types of analyses we want to do, and suggest paths forward for us all in the future
- Think about the nature of the data and the questions we want to answer, and *then* figure out the right metrics, rather than the other way around

- A note on distributions
- Why is linear least-squares regression inappropriate for (most) aerosol data analyses?
- What are the consequences of its misuse?
- What are some alternative useful metrics for aerosol data evaluation/comparison?
- Some other sticky problems

- A note on distributions
- Why is linear least-squares regression inappropriate for (most) aerosol data analyses?
- What are the consequences of its misuse?
- What are some alternative useful metrics for aerosol data evaluation/comparison?
- Some other sticky problems

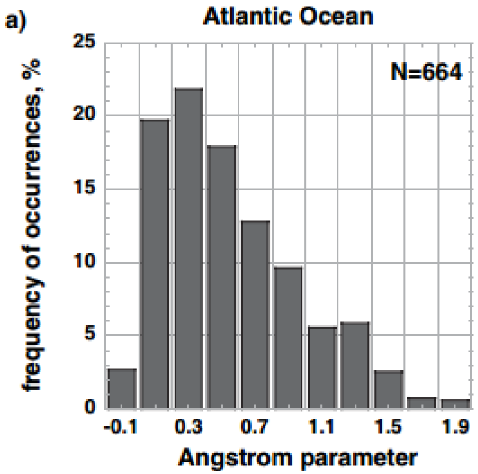
# Near-Lognormality of AOD means Gaussian statistics can be misleading

- Arithmetic means and standard deviations are poor representations of typical AOD and AOD variability
  - Long positive tail in AOD distributions
  - Implications for how comparisons and aggregates are done…
- Note doing linear regression in log-AOD space **does not** fix the problems in linear space

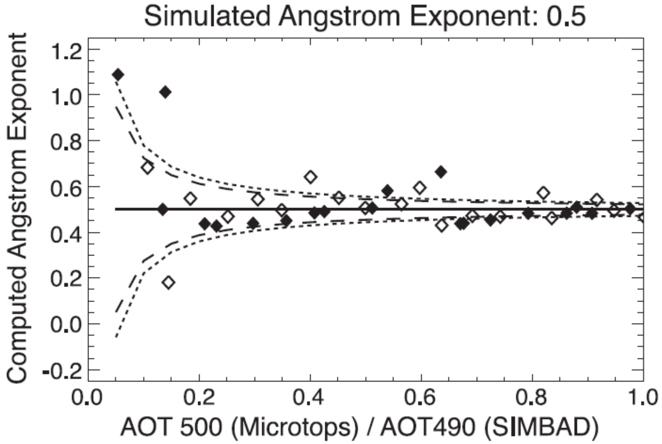


*O'Neill et al., GRL (2001), doi:10.1029/2000GL011581*

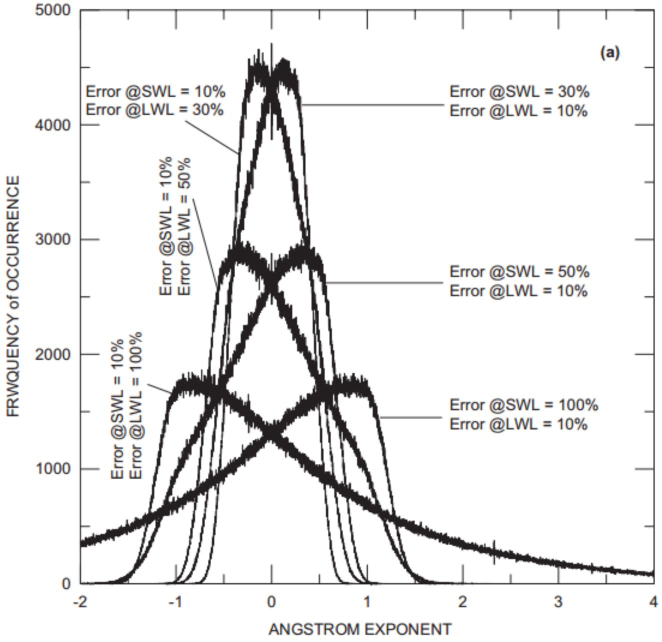
# Other quantities aren’t necessarily Gaussian, and may have AOD-dependent uncertainties



*Smirnov et al., AMT (2011), doi:10.5194/amt-4-583-2011*



*Knobelspiesse et al., RSE (2004), doi:10.1016/j.rse.2004.06.018*

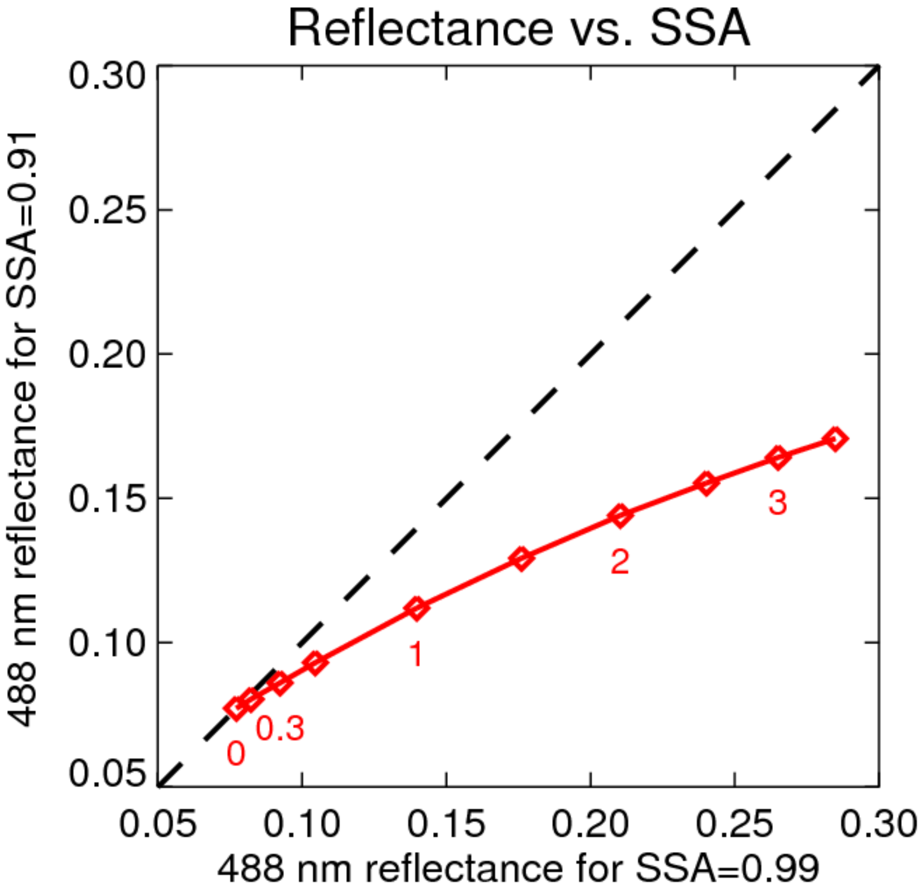


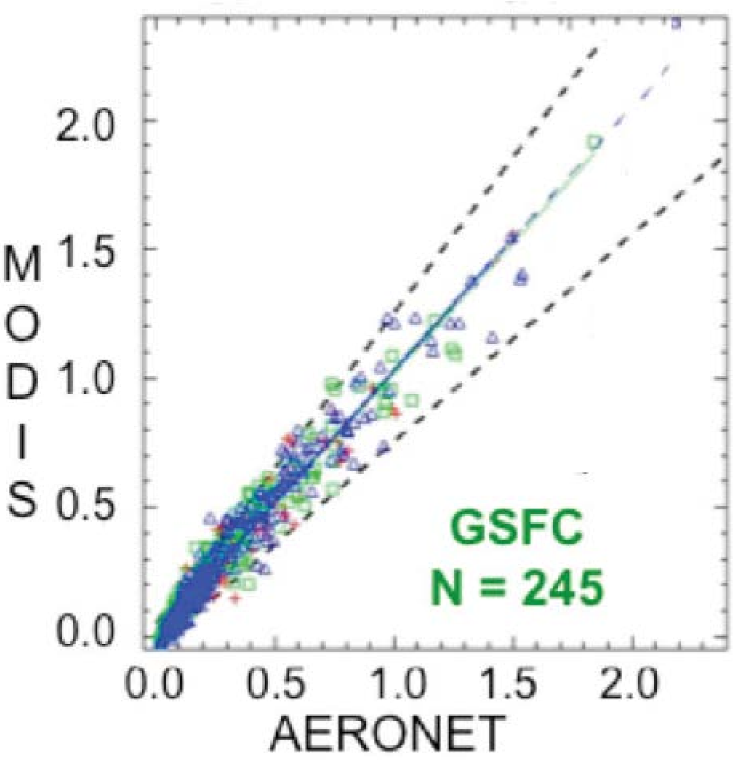
*Wagner and Silva, ACP (2008), doi:10.5194/acp-8-481-2008*

- Also has implications for data aggregation and sensitivity studies
- **Cannot** really validate them in low-AOD conditions
- See also fine mode fraction, SSA…

- A note on distributions
- Why is linear least-squares regression inappropriate for (most) aerosol data analyses?
- What are the consequences of its misuse?
- What are some alternative useful metrics for aerosol data evaluation/comparison?
- Some other sticky problems

# Assumption 1: linear relationship between quantities

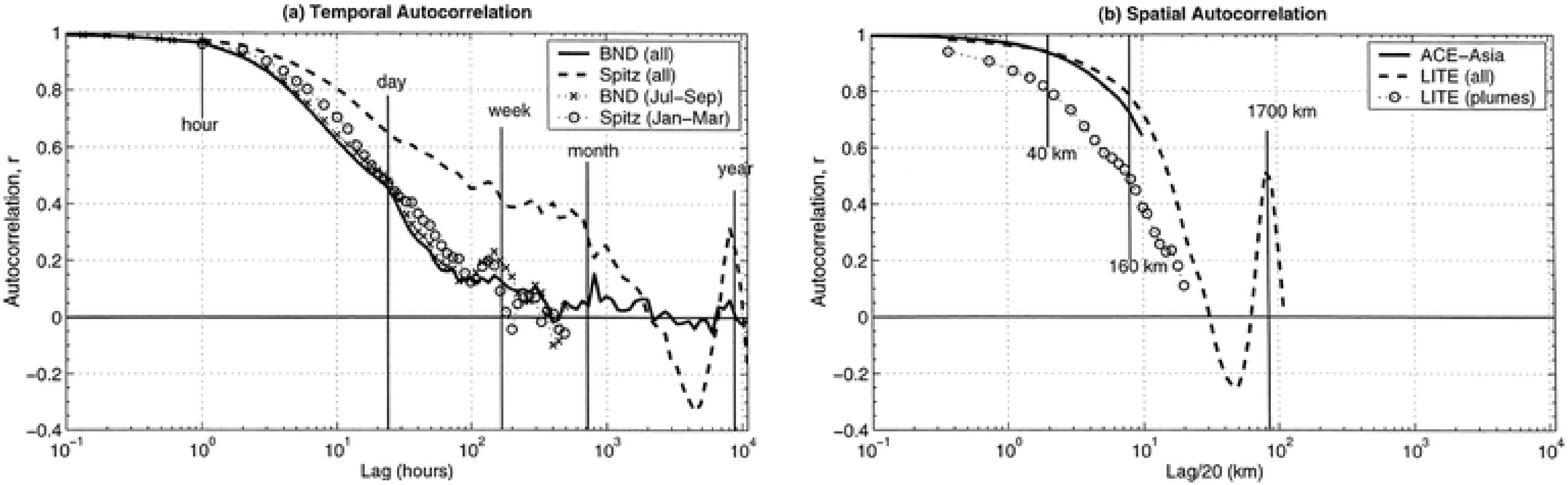




*Levy et al., ACP (2010), doi:10.5194/acp-10-10399-2010*

- Verdict: sometimes valid, not guaranteed (or expected)

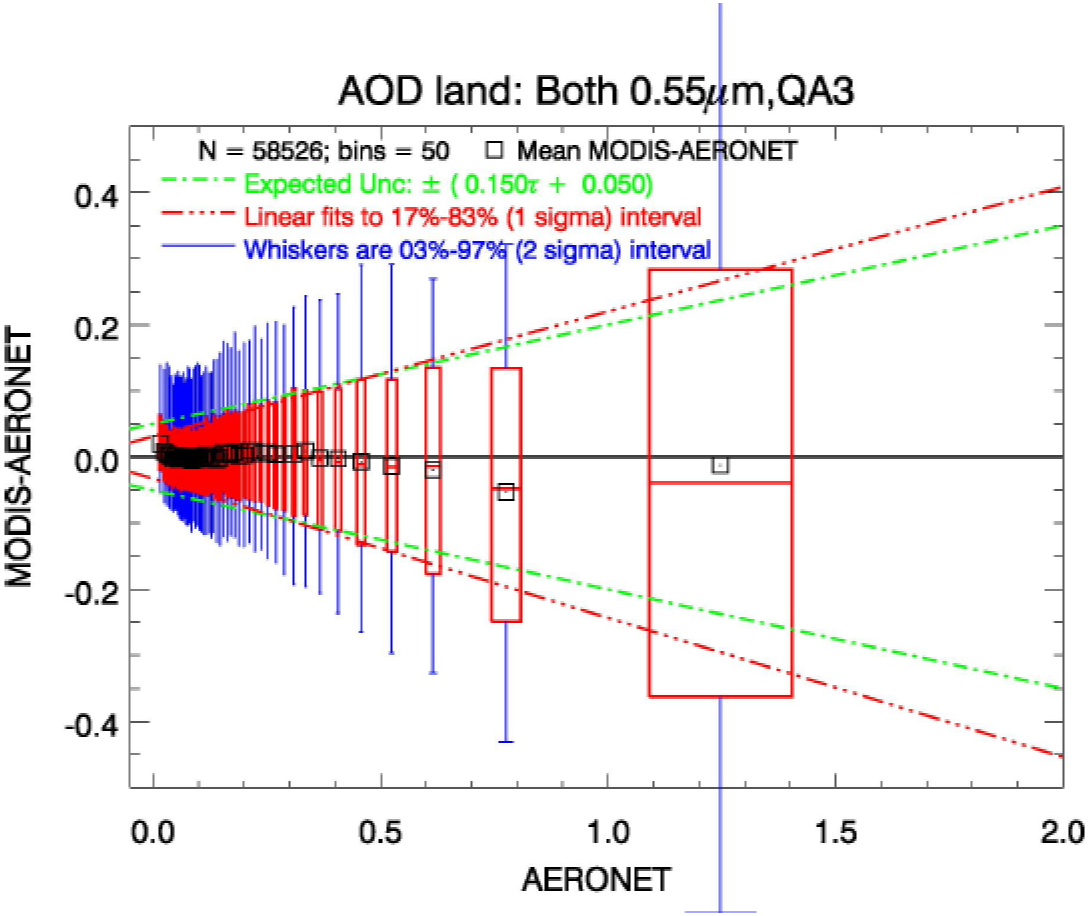
# Assumption 2: independence of data/errors



*Anderson et al., JAS (2003), doi:10.1175/1520-0469(2003)060<0119:MVOTA>2.0.CO;2*

- Verdict: invalid! Spatial and temporal autocorrelation.
- Also decreases the apparent variance in the data...

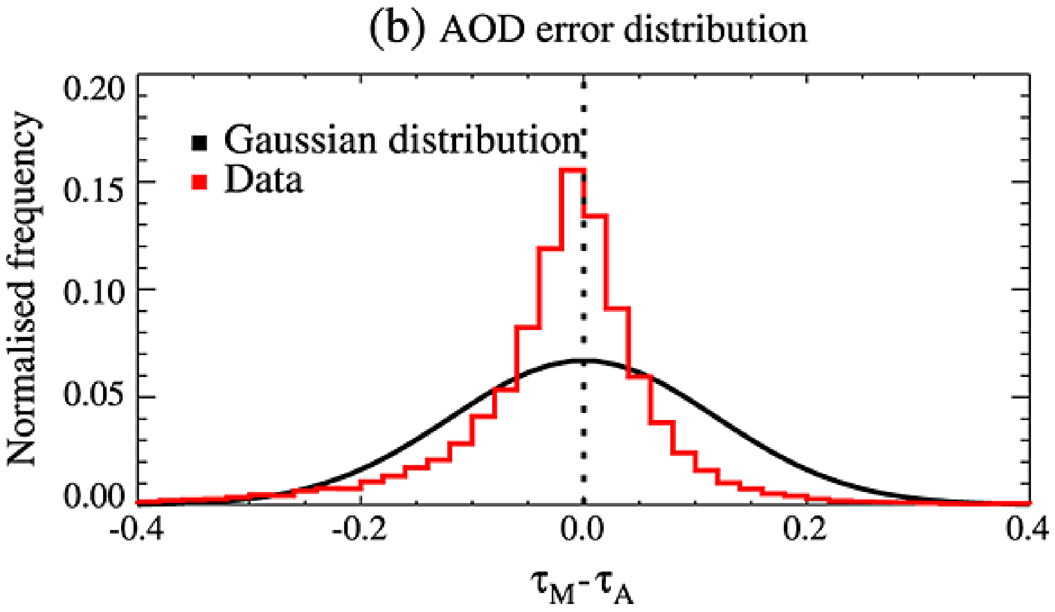
# Assumption 3: homoscedasticity (constant variance) of errors



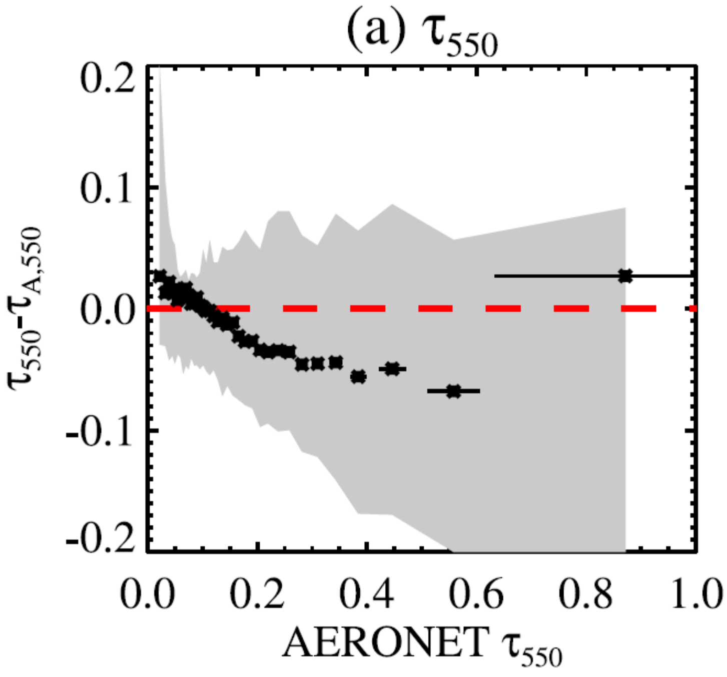
*Levy et al., ACP (2010), doi:10.5194/acp-10-10399-2010*

- Verdict: invalid! AOD uncertainty is AOD-dependent (among other things).

# Assumption 4: normality of errors



*Sayer et al., JGR (2013), doi:10.1002/jgrd.50600*

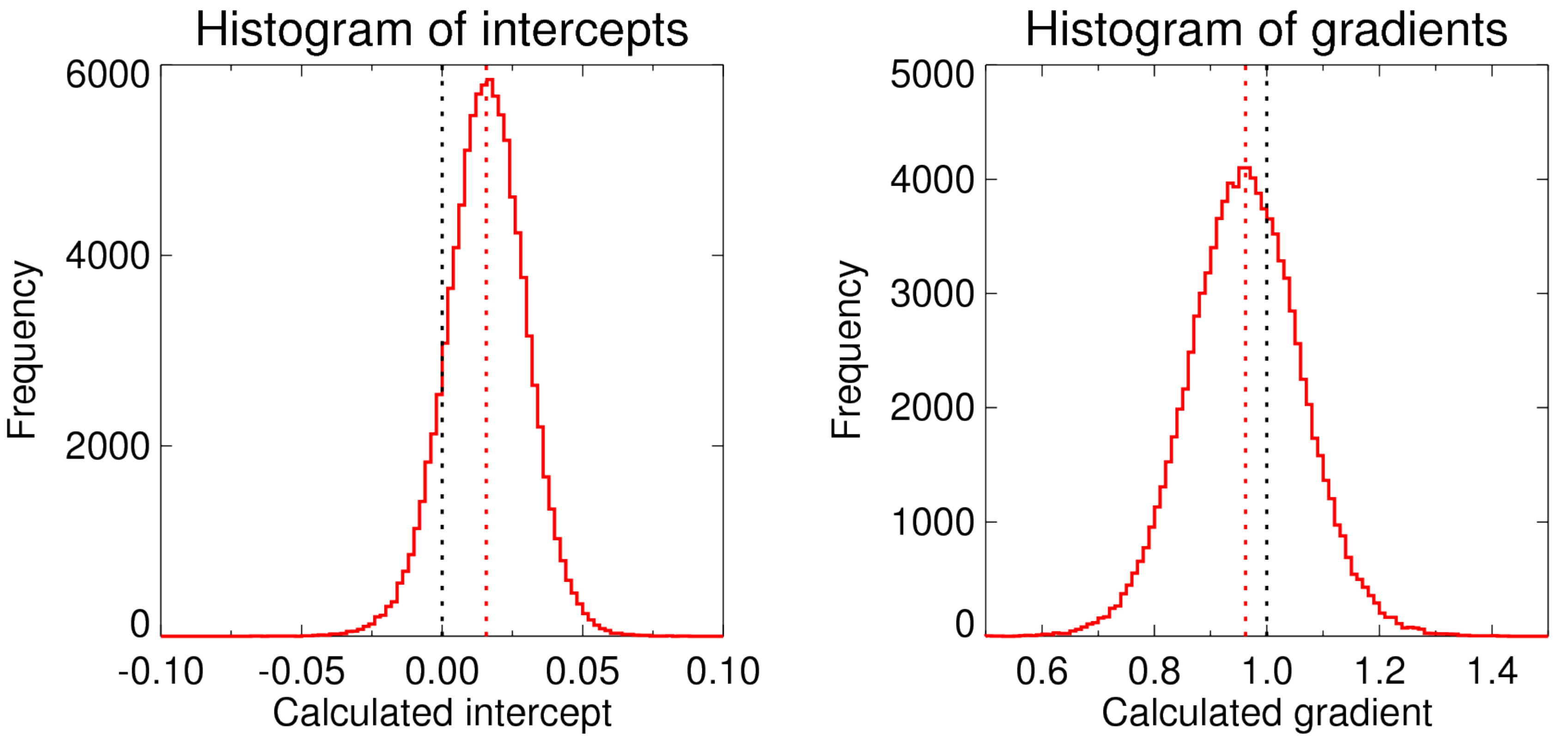


*Sayer et al., JGR (2012), doi:10.1029/2011JD016599*

- Verdict: invalid! Violations for both low-AOD and high-AOD conditions, for multiple reasons.

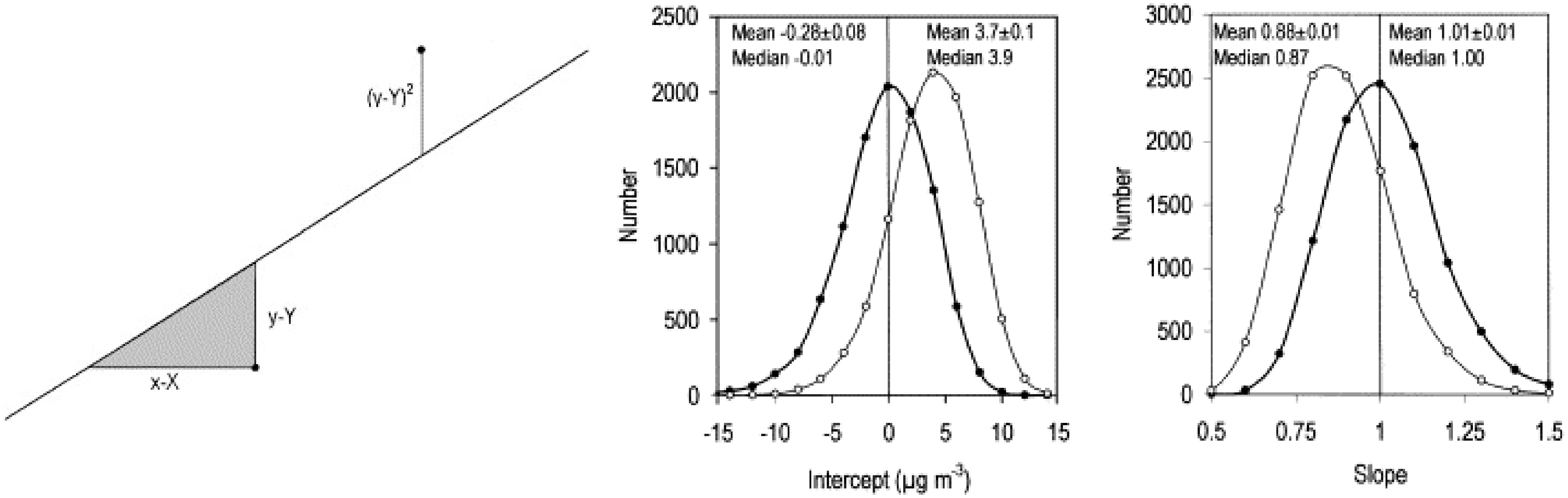
- A note on distributions
- Why is linear least-squares regression inappropriate for (most) aerosol data analyses?
- What are the consequences of its misuse?
- What are some alternative useful metrics for aerosol data evaluation/comparison?
- Some other sticky problems

# Regression output becomes biased thus misleading, even for an unbiased but noisy retrieval



- Intercept overestimated, slope underestimated due to error characteristics
  - $10^6$ runs of 100-member ensemble, 0.05+15% uncertainty, AOD lognormal ($-1,0.4^2$)
- Impact of linearity/independence/normality assumptions **harder to quantify**

# Reduced major axis (aka RMA, bivariate) fitting is a partial solution to the issue

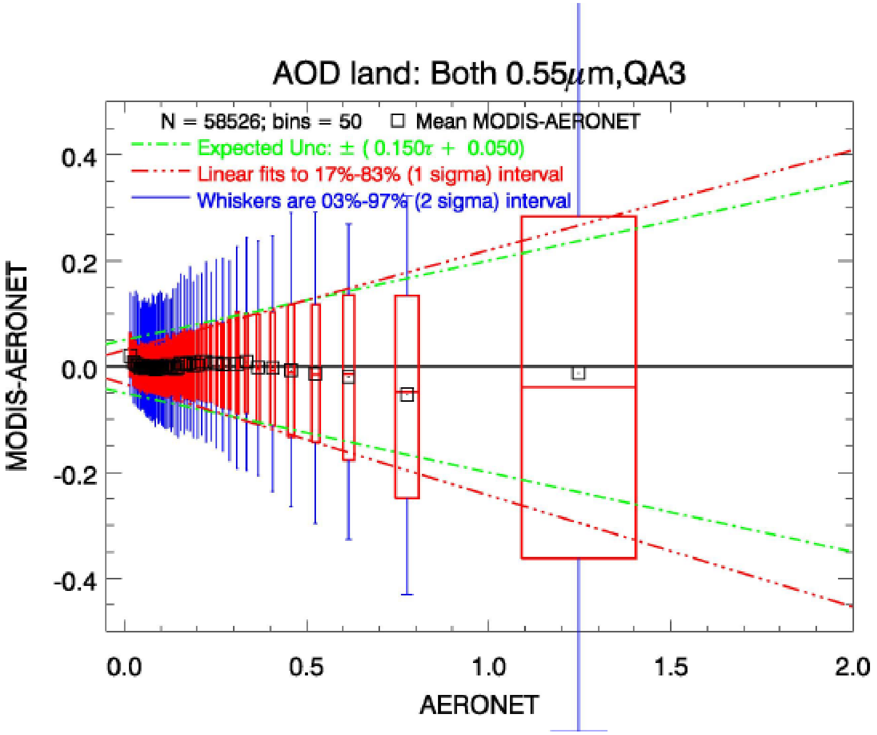


*Ayers, Atm. Env. (2001), doi:10.1016/S1352-2310(00)00527-6*

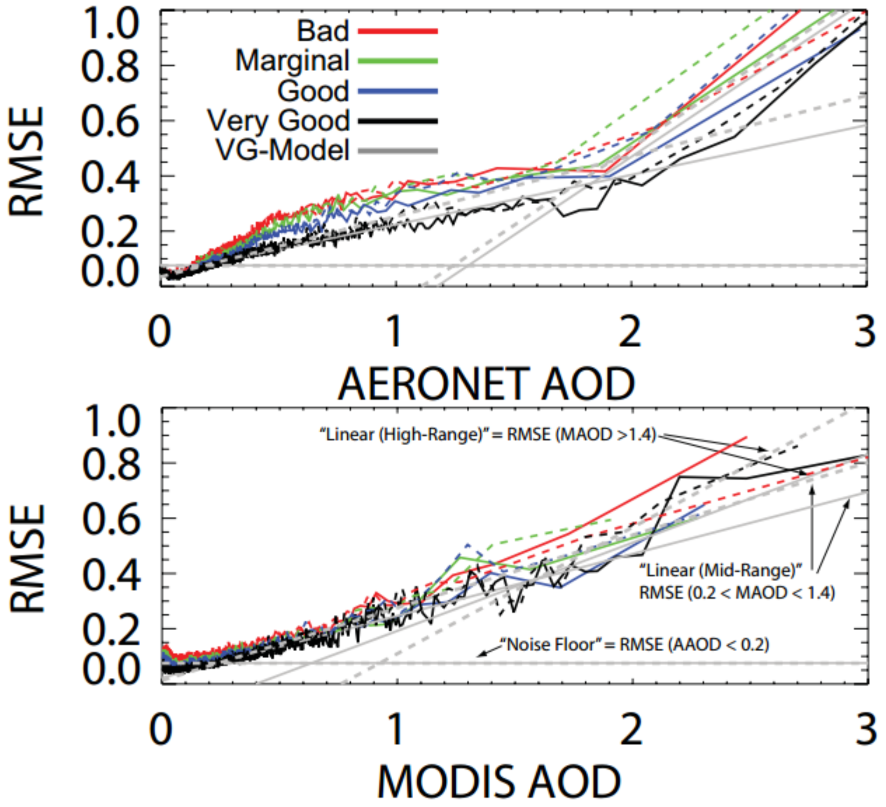
- Does **not** deal with linearity/independence/normality assumptions
- Can account for variable errors, and uncertainty in reference (true) data

- A note on distributions
- Why is linear least-squares regression inappropriate for (most) aerosol data analyses?
- What are the consequences of its misuse?
- What are some alternative useful metrics for aerosol data evaluation/comparison?
- Some other sticky problems

# Useful metric:
# error statistics vs. AOD



*Levy et al., ACP (2010),*
*doi:10.5194/acp-10-10399-2010*



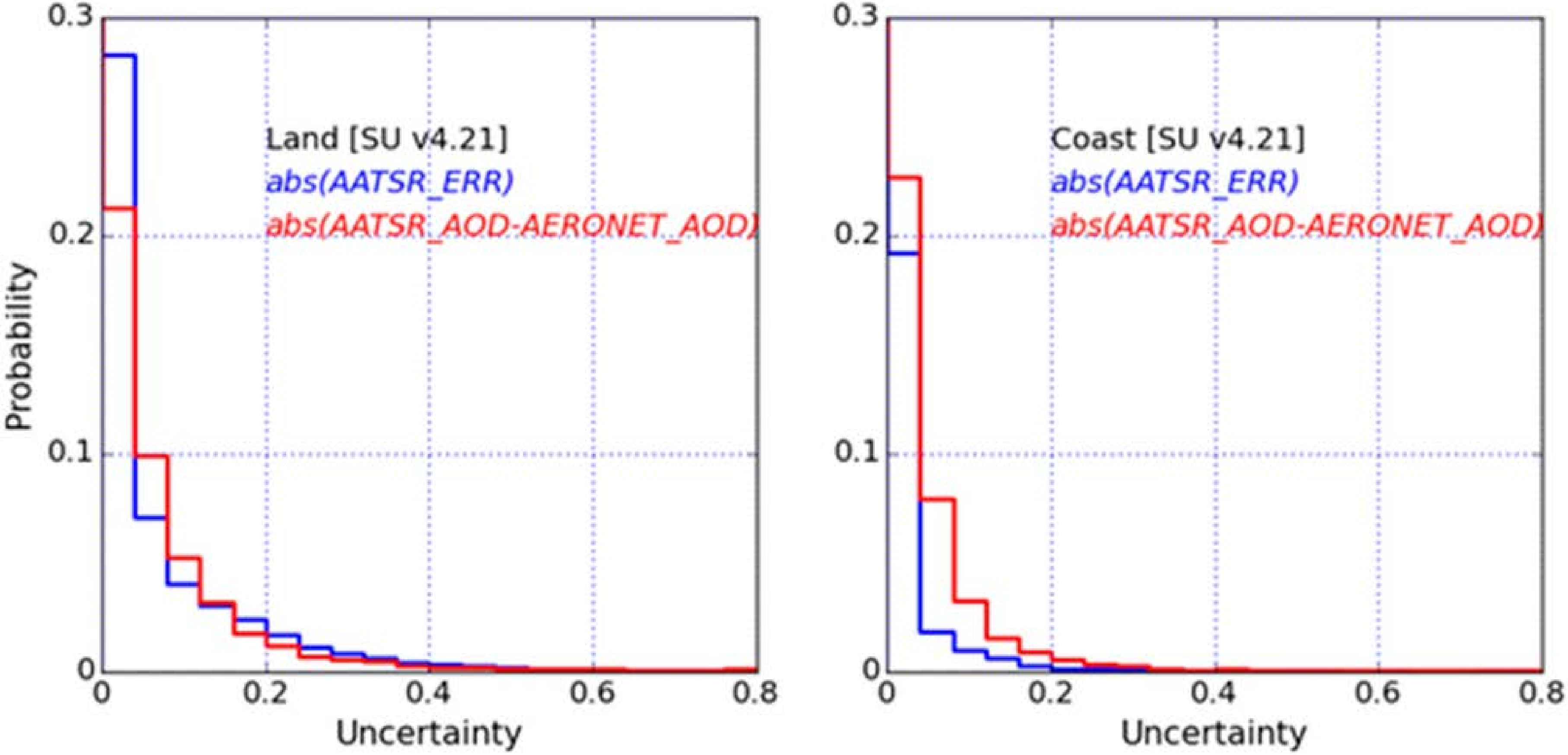
"Linear (High-Range)" = RMSE (MAOD >1.4)

"Linear (Mid-Range)" RMSE (0.2 < MAOD < 1.4)

"Noise Floor" = RMSE (AAOD < 0.2)

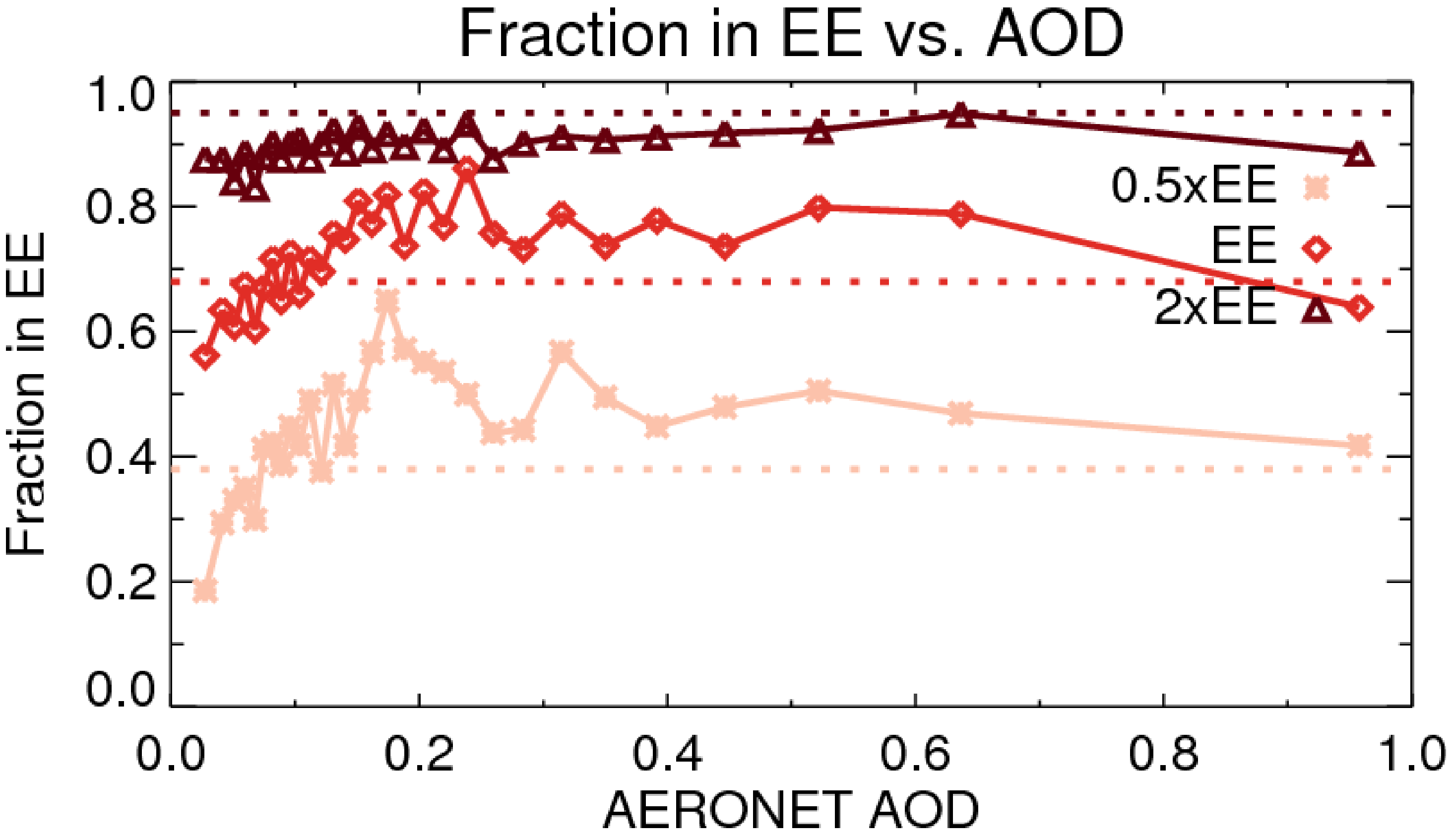
*Hyer et al., AMT (2011)*
*doi:10.5194/amt-4-379-2011*

# Useful metric:
# Compliance with uncertainty estimates



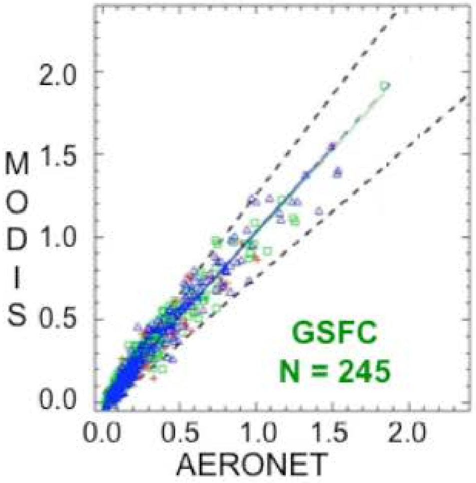
*Popp et al., Remote Sens. (2016), doi:10.3390/rs8050421*

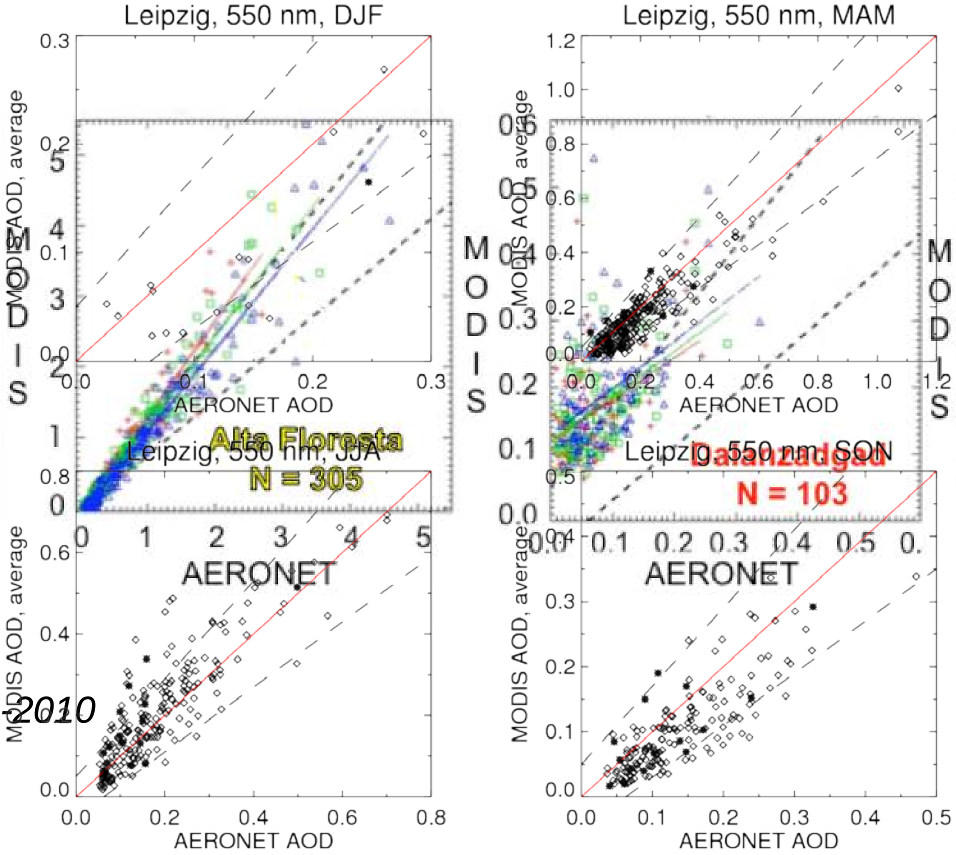
# Useful metric: Compliance with uncertainty estimates

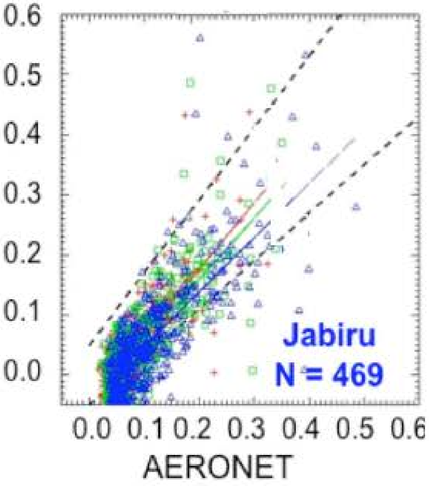


- A note on distributions
- Why is linear least-squares regression inappropriate for (most) aerosol data analyses?
- What are the consequences of its misuse?
- What are some alternative useful metrics for aerosol data evaluation/comparison?
- Some other sticky problems

# A more fundamental issue: what is our definition of the population?



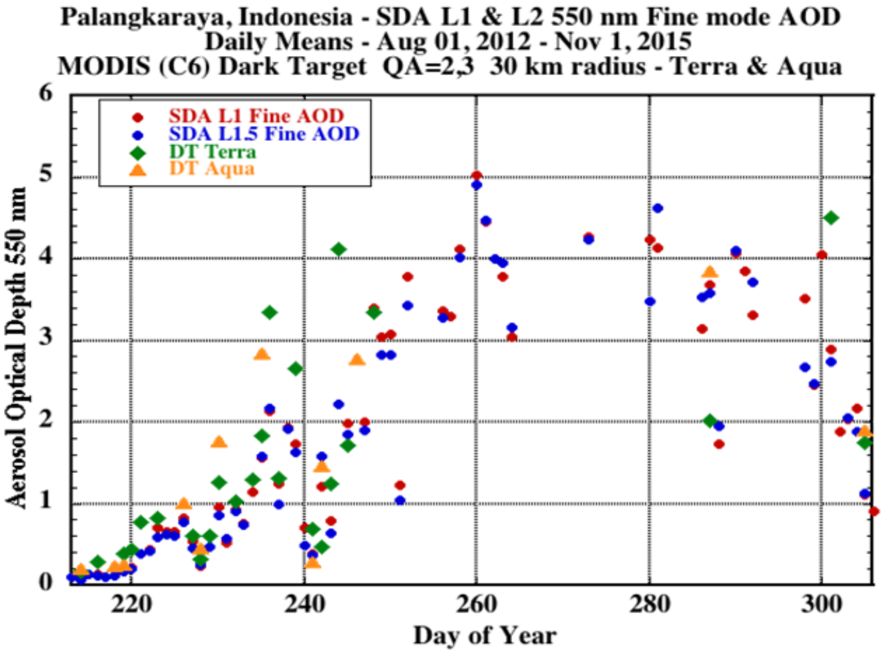


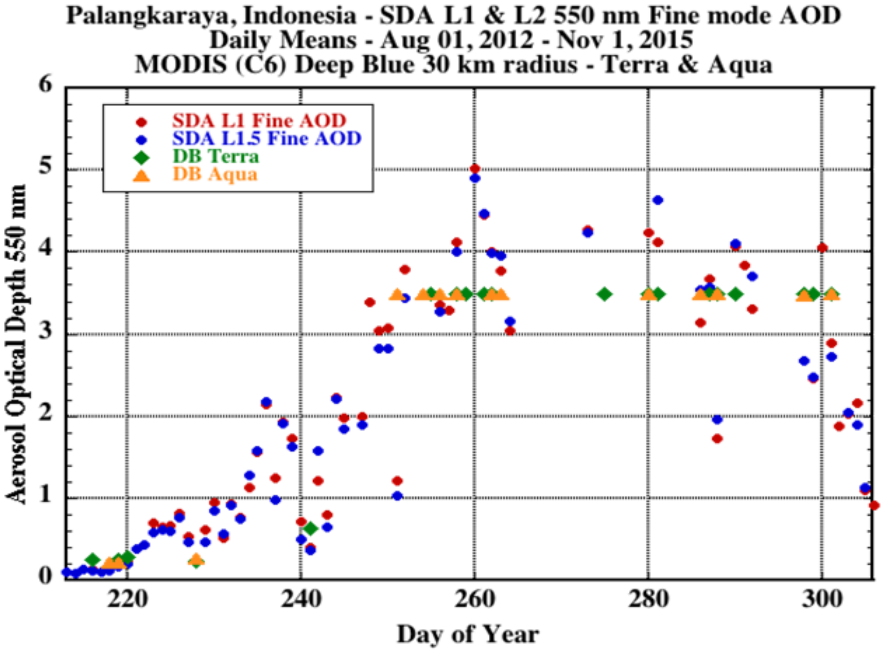


*Levy et al., ACP (2010), doi:10.5194/acp-10-10399-2010*

- Many statistical tests assume we are doing an analysis of samples drawn from **one** population
- Simple global aggregate statistics may not be meaningful for many analyses

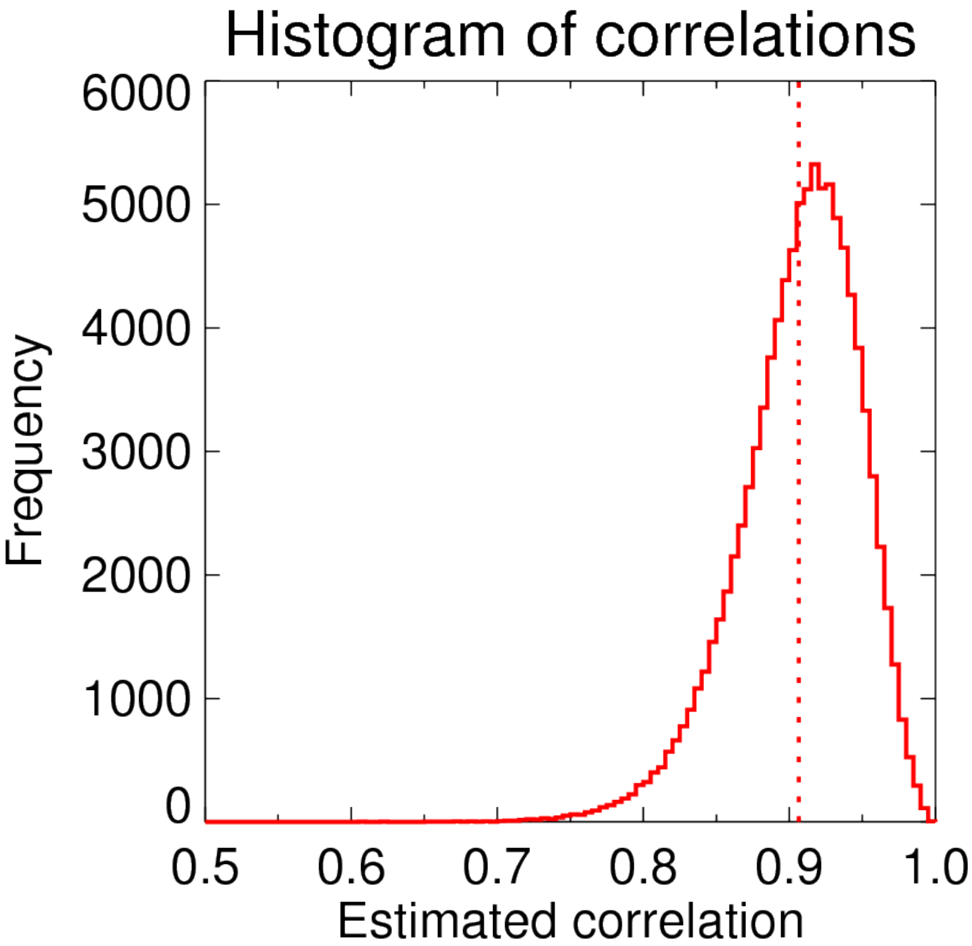
# Validation doesn't tell us about the events we miss





*Courtesy Tom Eck,*
*GESTAR-USRA/NASA GSFC*

# The sample statistics we calculate are only uncertain estimates of the population’s behaviour



# Other discussion points

- What do we want from validation/intercomparison exercises?
  - Uncertainties relative to ‘truth’?
  - Assess consistency between datasets?
  - Should location-based comparisons be the main focus when errors are mainly contextual?
- What are appropriate spatial/temporal scales for level 3 products?
  - What is Level 3 uncertainty?
- What do we want from correlation coefficients?
  - Should we use a rank correlation?
  - Estimate autocorrelation?
- How should we treat AERONET variability and uncertainty?
  - Legitimate sampling differences can appear as outliers
  - Gaussian vs. lognormal statistics
- What should we spend more time looking at?
  - Defining ‘events’ and frequency of their omission?
  - Retrieval coverage?
- What about Ångström exponent and single scattering albedo?
  - Some of the same issues, some different characteristics…

# Some useful resources

- Wikipedia pages:
  - Summary on regression analysis https://en.wikipedia.org/wiki/Regression_analysis
  - Linear regression https://en.wikipedia.org/wiki/Linear_regression
  - Pearson correlation coefficient https://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient
  - Rank correlation overview https://en.wikipedia.org/wiki/Rank_correlation
- Prof. Nau's (Duke) webpages on linear regression http://people.duke.edu/~rnau/testing.htm
- Wolfram Mathworld page on linear regression http://mathworld.wolfram.com/LeastSquaresFitting.html
- Schönbrodt & Perugini (2013), At what sample size do correlations stabilize? *Journal of Research in Personality*, 47, 609-612, doi:10.1016/j.jrp.2013.05.009