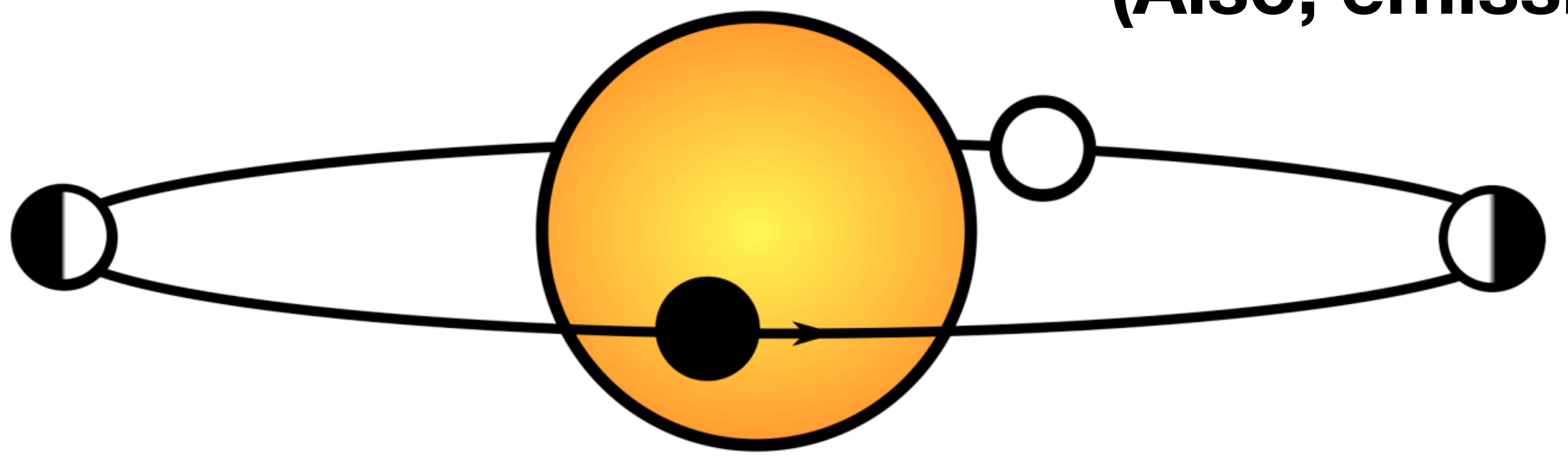
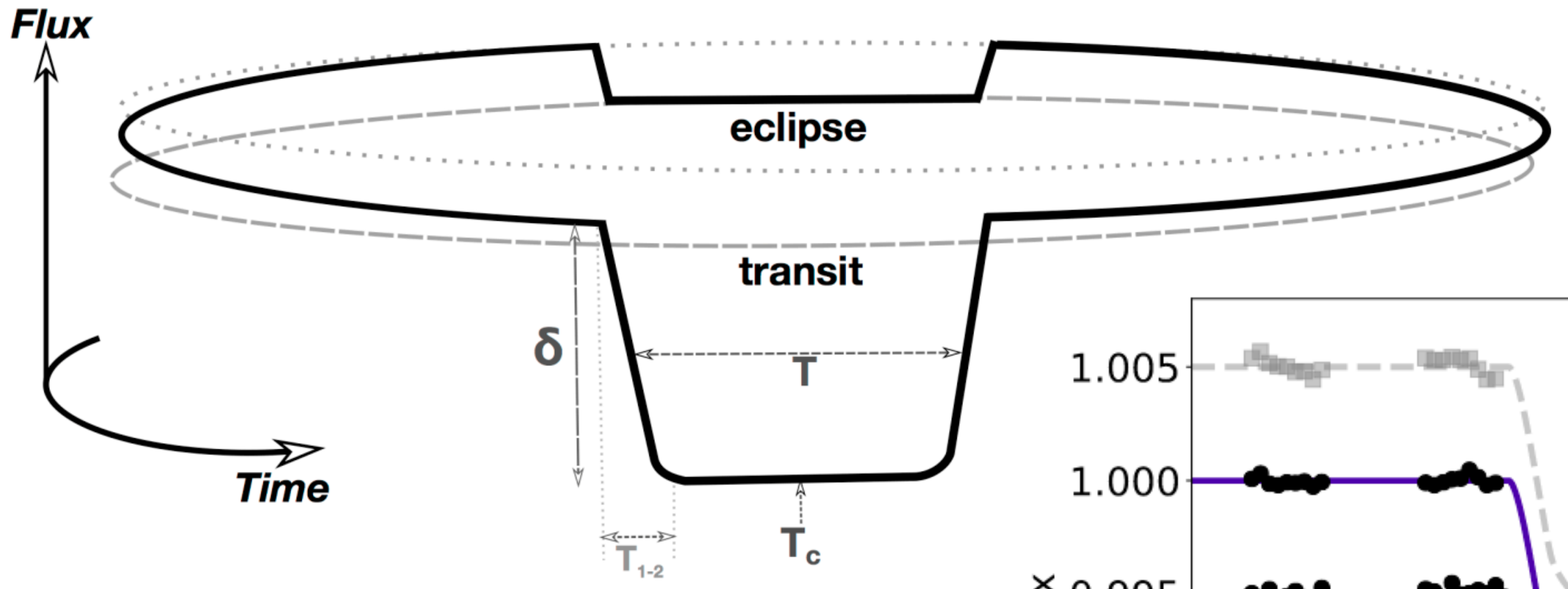
The Pitfalls of Bayes: On the Use of Statistical Goodness of Fit Criteria in the Evaluation of Transmission Spectra

Tom J Wilson
onoddil@pm.me
University of Exeter

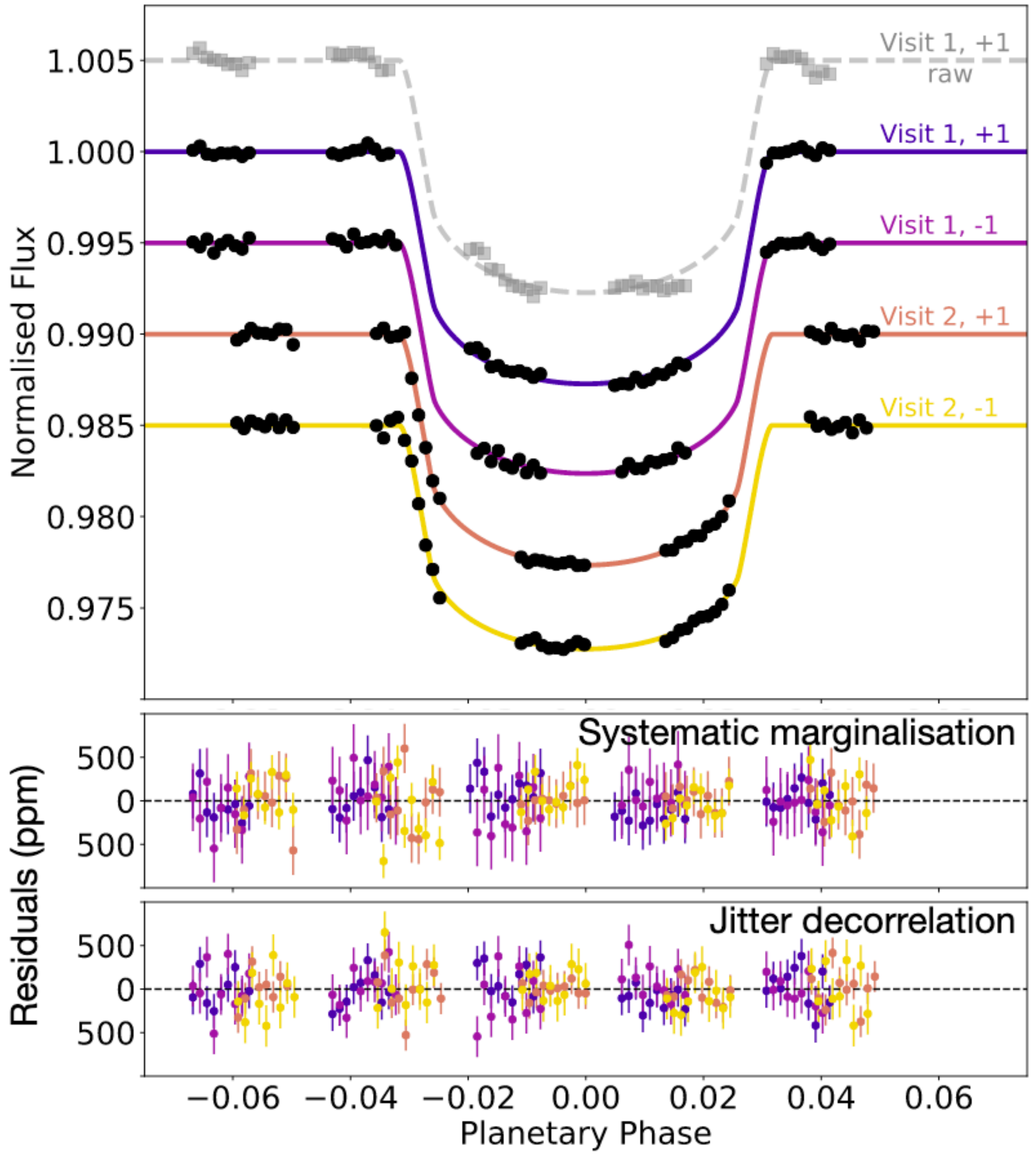
Exoplanet Transmission Spectra

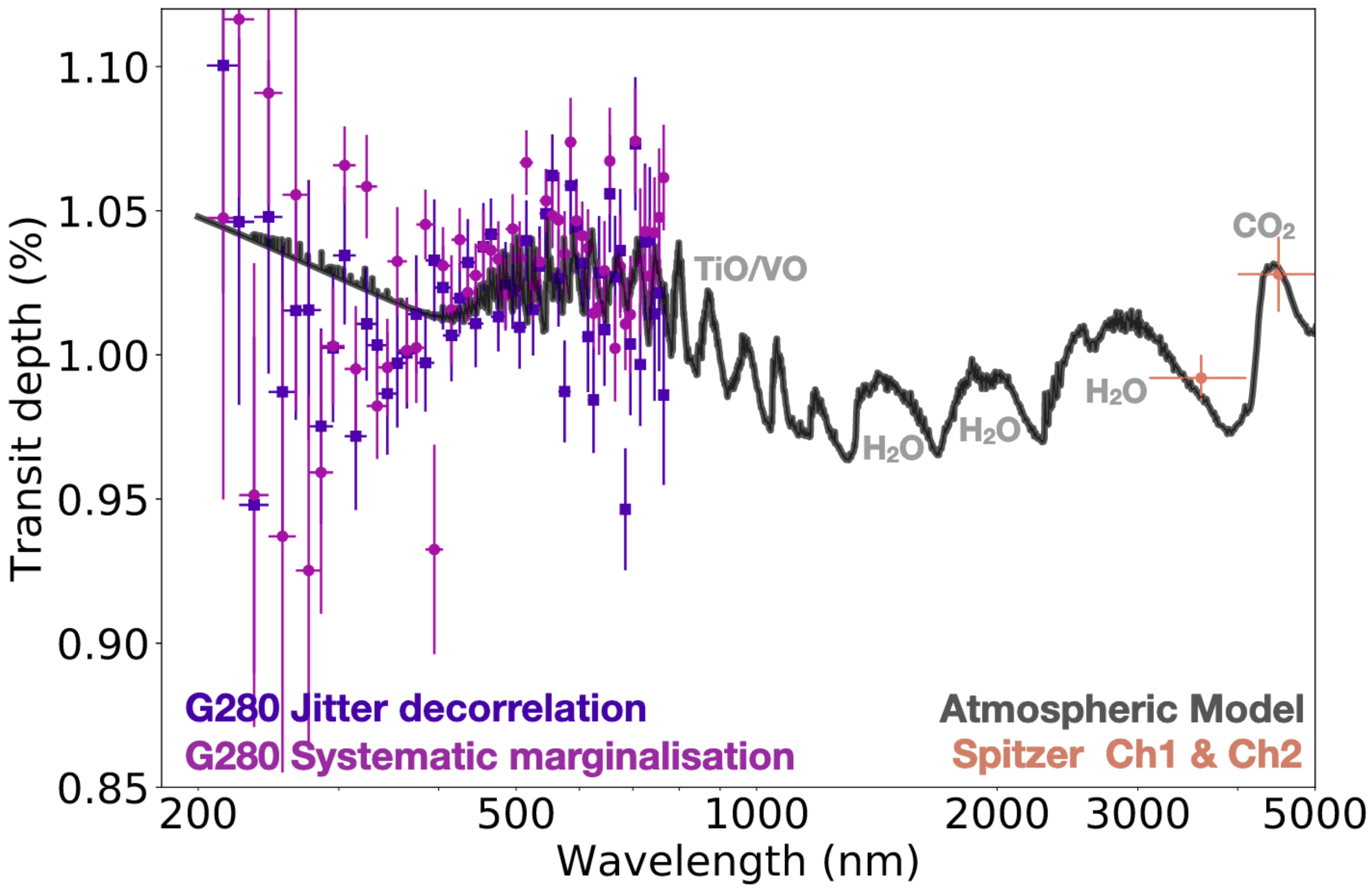
(Also, emission spectra!)





H. Wakeford, stellarplanet.org





Wakeford et al. (2020)

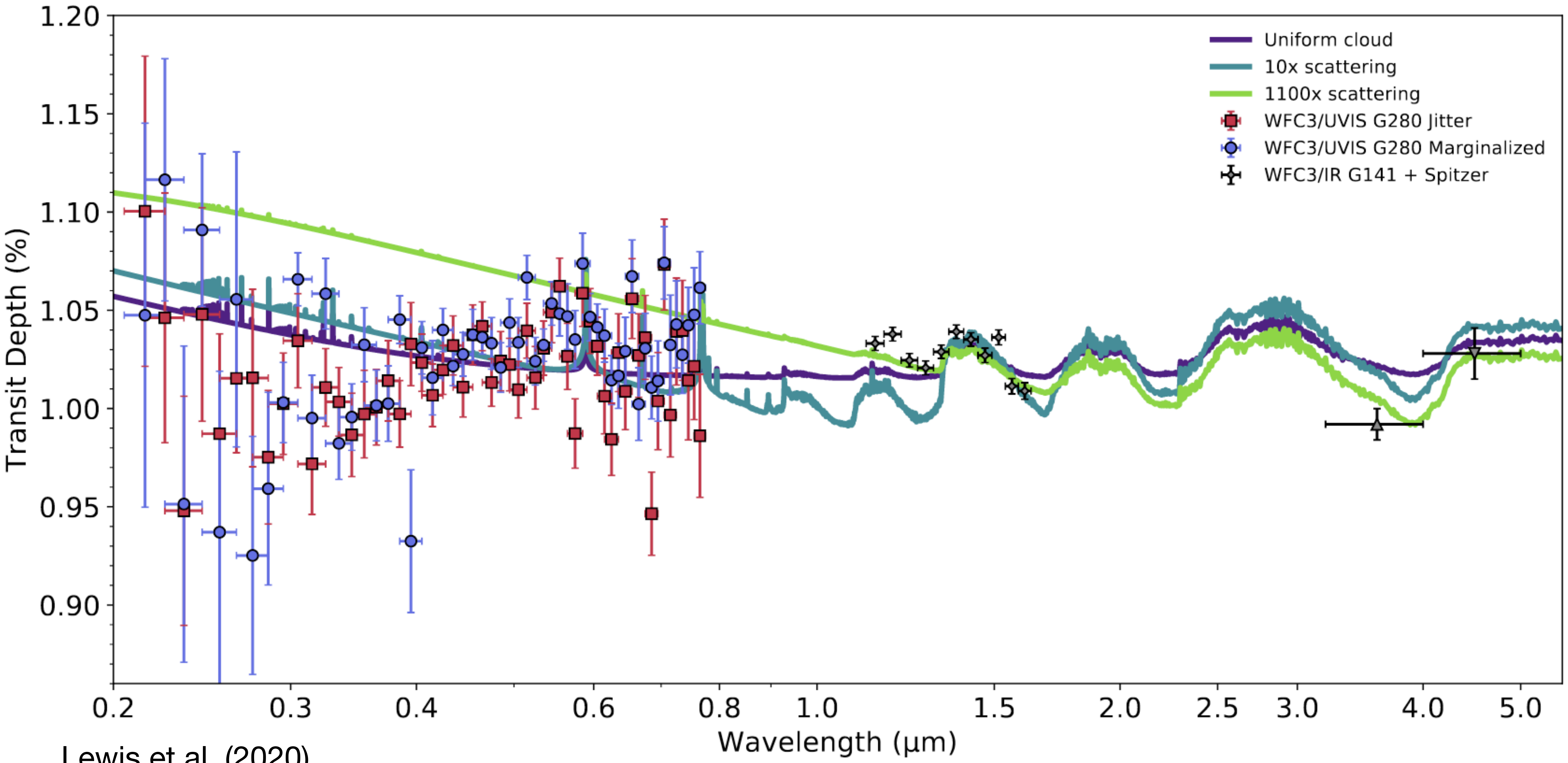
Tom J Wilson @onoddil

What's in an Exoplanet Atmosphere?

A definitely non-exhaustive list of model fitting methods:

- **1-D models**
- **3-D models, GCMs**

1. Equilibrium
2. Disequilibrium
A. Chemistry models
B. Clouds/Hazes
❋ Self-consistent models
❋ Parametric models
◆ GPs
◦ Forward modelling
  ◦ Grid search
◦ Retrievals
  ◦ Nested sampling
  ◦ MCMC, Monte Carlo



Lewis et al. (2020)

Tom J Wilson @onoddil

What’s in an Exoplanet Atmosphere?

A definitely non-exhaustive list of model fitting methods:

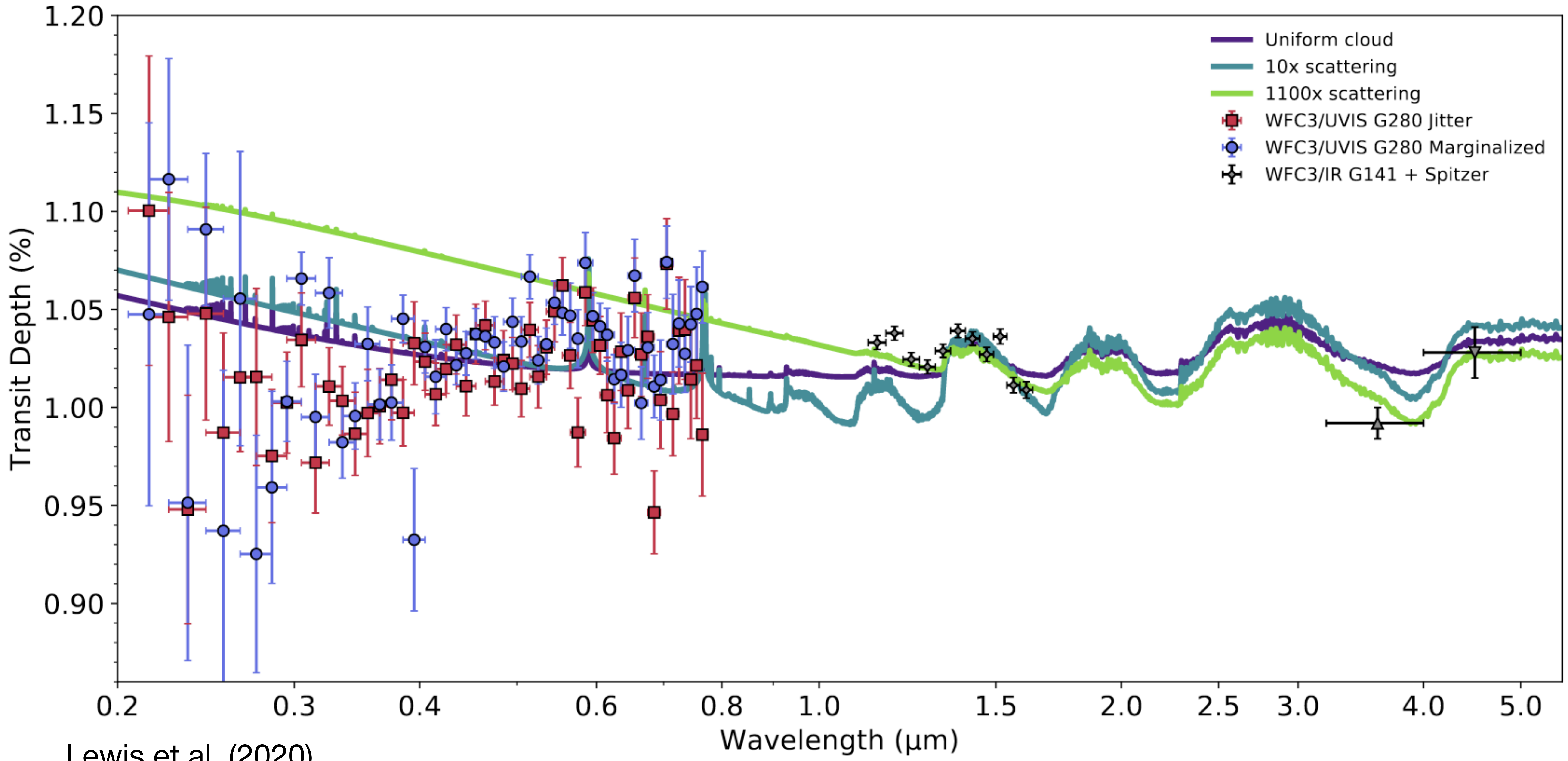
- **1-D models**
- **3-D models, GCMs**

1. Equilibrium
2. Disequilibrium

A. Chemistry models
B. Clouds/Hazes

❋ Self-consistent models
❋ Parametric models
◆ GPs

- **Forward modelling**
  - **Grid search**
- **Retrievals**
  - **Nested sampling**
  - **MCMC, Monte Carlo**



Lewis et al. (2020)

What's the best model fit to the data?

(And what do "best" and "model" mean anyway?)

Model in a *suite* of models (full chemistry, no H_2O, no CH_4, etc.)

Parameterisation of a *given* model (T-P profile, $\log(X_{H2O})$, radius, etc.)

Likelihood

Prior

$$p\left(\boldsymbol{\theta} | \boldsymbol{y}_{\text{obs}}, M_i\right) \equiv \frac{\mathcal{L}\left(\boldsymbol{y}_{\text{obs}} | \boldsymbol{\theta}, M_i\right) \pi\left(\boldsymbol{\theta} | M_i\right)}{\mathcal{Z}\left(\boldsymbol{y}_{\text{obs}} | M_i\right)}$$

Evidence

$$\mathcal{Z}\left(\boldsymbol{y}_{\text{obs}} | M_i\right) = \int_{\text{all}\,\boldsymbol{\theta}} \mathcal{L}\left(\boldsymbol{y}_{\text{obs}} | \boldsymbol{\theta}, M_i\right) \pi\left(\boldsymbol{\theta} | M_i\right) \mathrm{d}\boldsymbol{\theta}$$

MacDonald & Madhusudhan (2017)

Evidence either comes from the sampled posterior (e.g. nested sampling, MCMC) or can be derived from the maximum likelihood (e.g. forward models, grid search) through the BIC and AIC: Bayesian/Akaike Information Criterion

Number of data points

$$\mathrm{BIC} \equiv k \ln(n) - 2 \ln(\hat{L})$$

$$\mathrm{AIC} \equiv 2k - 2 \ln(\hat{L})$$

Number of parameters

Maximum value of the likelihood function

Maximum evidence is used to select the best model (from a *suite of models)* quite often in exoplanet characterisation literature

Tom J Wilson @onoddil

The Need for "Goodness of Fit"

The (log-)evidence provides a *relative* ranking of each model in a given suite; a Bayesian *classifier* as opposed to a *probability model*.

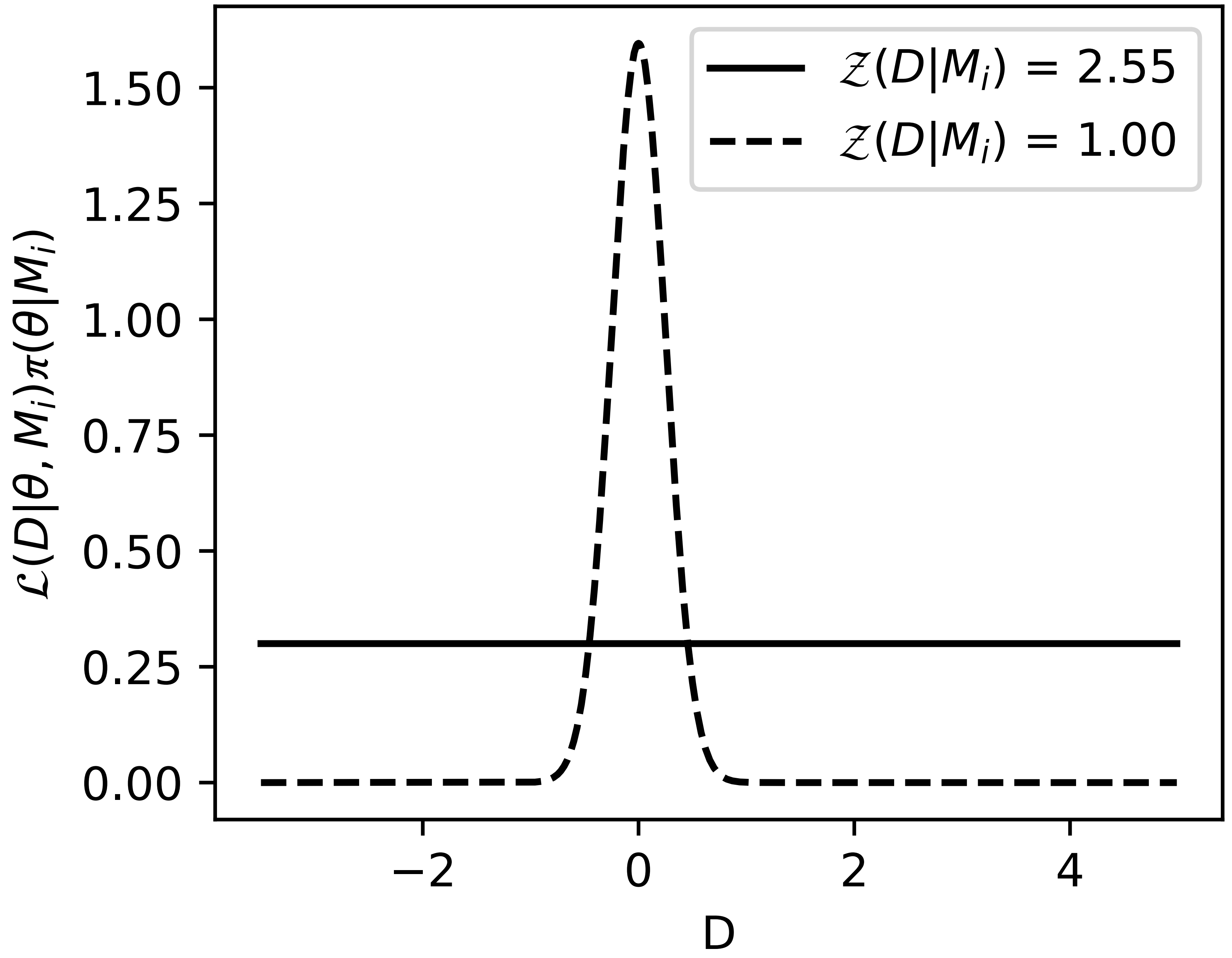
Bayes factors compare relative evidence of two models:

$$B_{01} = \frac{Z(D\,|\,H_0)}{Z(D\,|\,H_1)}$$ **(neglecting priors)**

E.g. Trotta (2008), Benneke & Seager (2013), after Kass & Rafferty (1995), again after Jeffreys (1961)

- **+ Easy to compute**
- **+ Not sensitive to exact parameters**
- **+ Obvious to interpret**
- **- No absolute grounding**
- **- Does not inform whether *any* model explains the data**

Important that the *posterior probability* is also computed to fully interpret the data fit.



Which is preferable: model with higher evidence, or one with lower evidence but with a better fit of a single parameterisation?

The Chi-Squared Statistic

Ignoring priors, the (log-)likelihood can be expressed as the chi-squared statistic.

$$Q = \sum_{i}^{n} \frac{(f(x_i) - y_i)^2}{\sigma_i^2} \sim \chi^2(n - k)$$

Degrees of freedom (DoF), ν

The *null hypothesis*, H_0, describes the idea that chance alone is responsible for one's results — in this case, the normalised residuals to the fit.

The Chi-Squared Statistic

Ignoring priors, the (log-)likelihood can be expressed as the chi-squared statistic.

$$Q = \sum_{i}^{n} \frac{(f(x_i) - y_i)^2}{\sigma_i^2} \sim \chi^2(n-k)$$

Degrees of freedom (DoF), ν

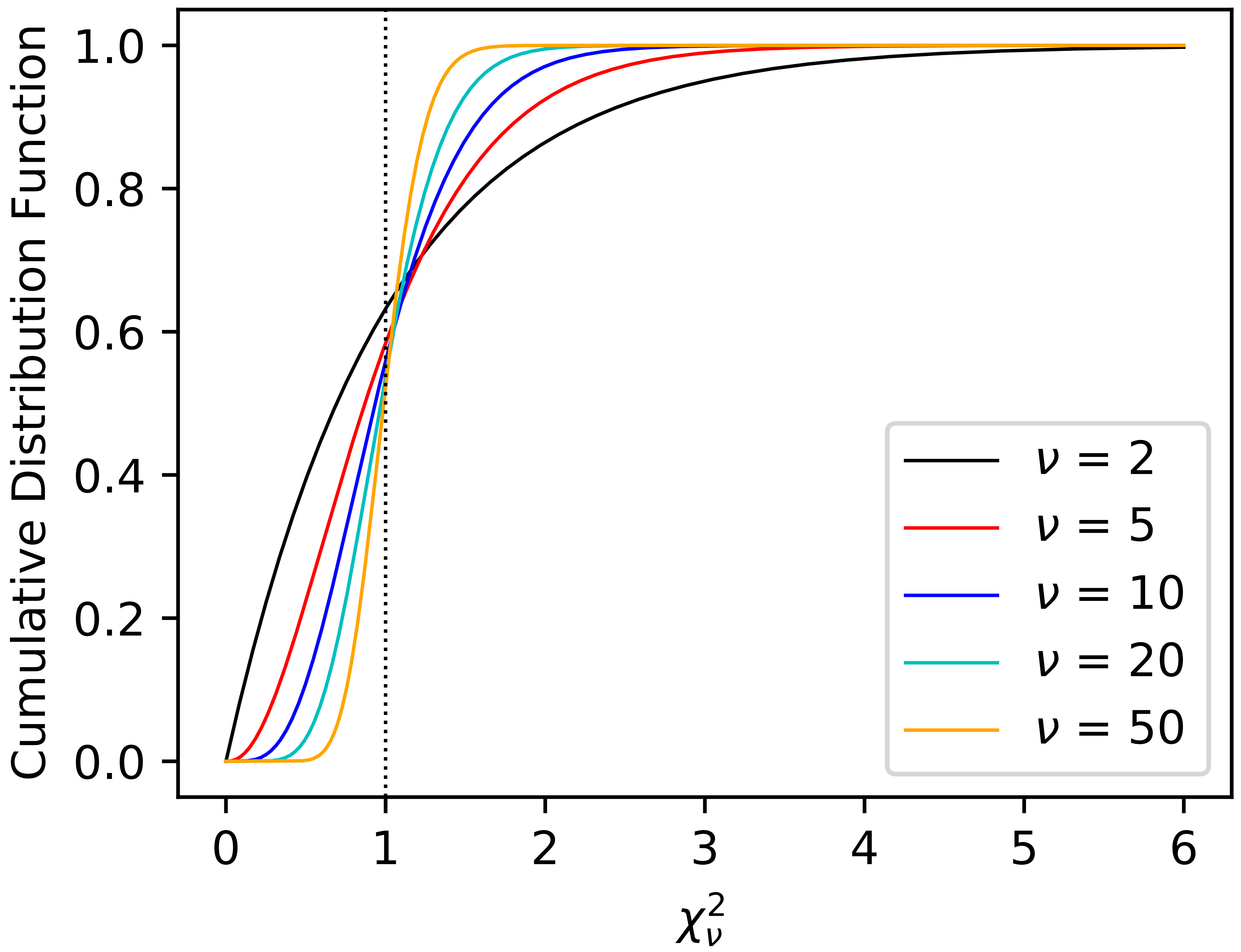
The *null hypothesis*, H_0, describes the idea that chance alone is responsible for one's results — in this case, the normalised residuals to the fit.

We can ask what the probability of getting a (reduced) chi-squared value equal or smaller than χ^2_ν purely by chance is, given by the cumulative integral of the chi-squared distribution:

Lower incomplete gamma function

$$P(\chi^2 \leq x) = F(x; k) = \frac{\gamma(\nu/2,\ x/2)}{\Gamma(\nu/2)}$$

Gamma function



Tom J Wilson @onoddil

The Chi-Squared Statistic

Ignoring priors, the (log-)likelihood can be expressed as the chi-squared statistic.

$$Q = \sum_i^n \frac{(f(x_i) - y_i)^2}{\sigma_i^2} \sim \chi^2(n-k)$$

Degrees of freedom (DoF), ν

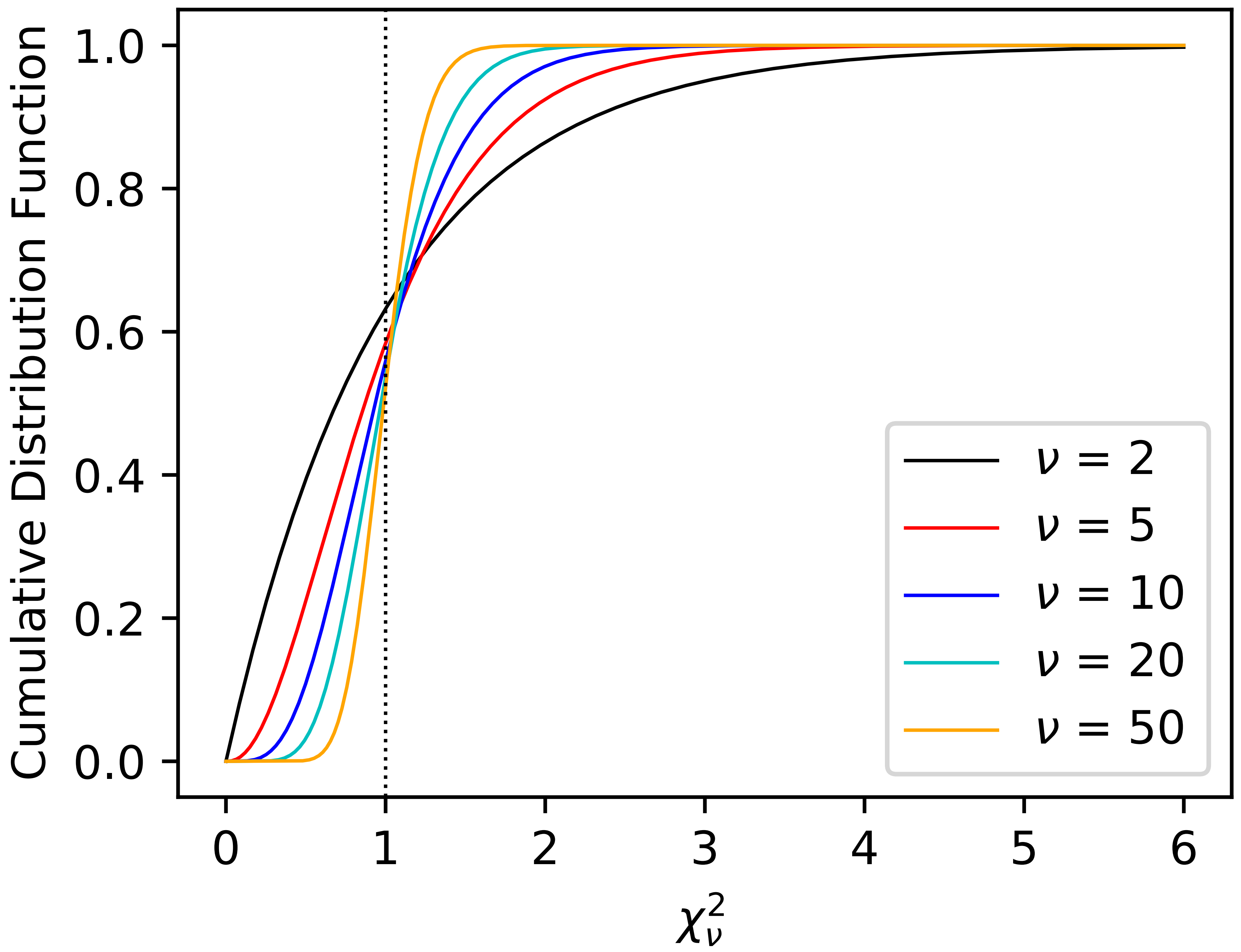
The *null hypothesis*, H_0, describes the idea that chance alone is responsible for one's results — in this case, the normalised residuals to the fit.

We can ask what the probability of getting a (reduced) chi-squared value equal or smaller than χ^2_ν purely by chance is, given by the cumulative integral of the chi-squared distribution:

Lower incomplete gamma function

$$P(\chi^2 \leq x) = F(x;k) = \frac{\gamma(\nu/2,\ x/2)}{\Gamma(\nu/2)}$$

Gamma function



Note that as ν increases, the acceptable confidence interval around "reduced chi-squared of one" decreases!

Always quote the (reduced) chi-squared AND degrees of freedom, and consider the rejection of the null hypothesis

Tom J Wilson @onoddil

The Bayesian Posterior Probability

Evidence-based posterior

E.g. Gibson (2014)

$$p(M_i\,|\,D) = \frac{Z(D\,|\,M_i)\,\pi(M_i)}{\sum_i Z(D\,|\,M_i)\,\pi(M_i)}$$

Maximum likelihood-based posterior

$$p(M_i\,|\,D, \hat{\theta}_i) = \frac{p(D\,|\,\hat{\theta}_i, M_i)\,p(\hat{\theta}_i\,|\,M_i)\pi(M_i)}{\sum_i p(D\,|\,\hat{\theta}_i, M_i)\,p(\hat{\theta}_i\,|\,M_i)\pi(M_i)}$$

The Bayesian Posterior Probability

Evidence-based posterior

E.g. Gibson (2014)

$$p(M_i\,|\,D) = \frac{Z(D\,|\,M_i)\,\pi(M_i)}{\sum_i Z(D\,|\,M_i)\,\pi(M_i)}$$

Maximum likelihood-based posterior

$$p(M_i\,|\,D,\hat{\theta}_i) = \frac{p(D\,|\,\hat{\theta}_i, M_i)\,p(\hat{\theta}_i\,|\,M_i)\pi(M_i)}{\sum_i p(D\,|\,\hat{\theta}_i, M_i)\,p(\hat{\theta}_i\,|\,M_i)\pi(M_i)}$$

The “null hypothesis” could be treated as a “fire extinguisher”, “to be held in abeyance until needed”.

Jaynes (2003), “Probability Theory: The Logic of Science"

The Bayesian Posterior Probability

Evidence-based posterior

E.g. Gibson (2014)

$$p(M_i\,|\,D) = \frac{Z(D\,|\,M_i)\,\pi(M_i)}{\sum_i Z(D\,|\,M_i)\,\pi(M_i)}$$

Maximum likelihood-based posterior

$$p(M_i\,|\,D, \hat{\theta}_i) = \frac{p(D\,|\,\hat{\theta}_i, M_i)\,p(\hat{\theta}_i\,|\,M_i)\pi(M_i)}{\sum_i p(D\,|\,\hat{\theta}_i, M_i)\,p(\hat{\theta}_i\,|\,M_i)\pi(M_i)}$$

The “null hypothesis” could be treated as a “fire extinguisher”, “to be held in abeyance until needed”.

Jaynes (2003), “Probability Theory: The Logic of Science"

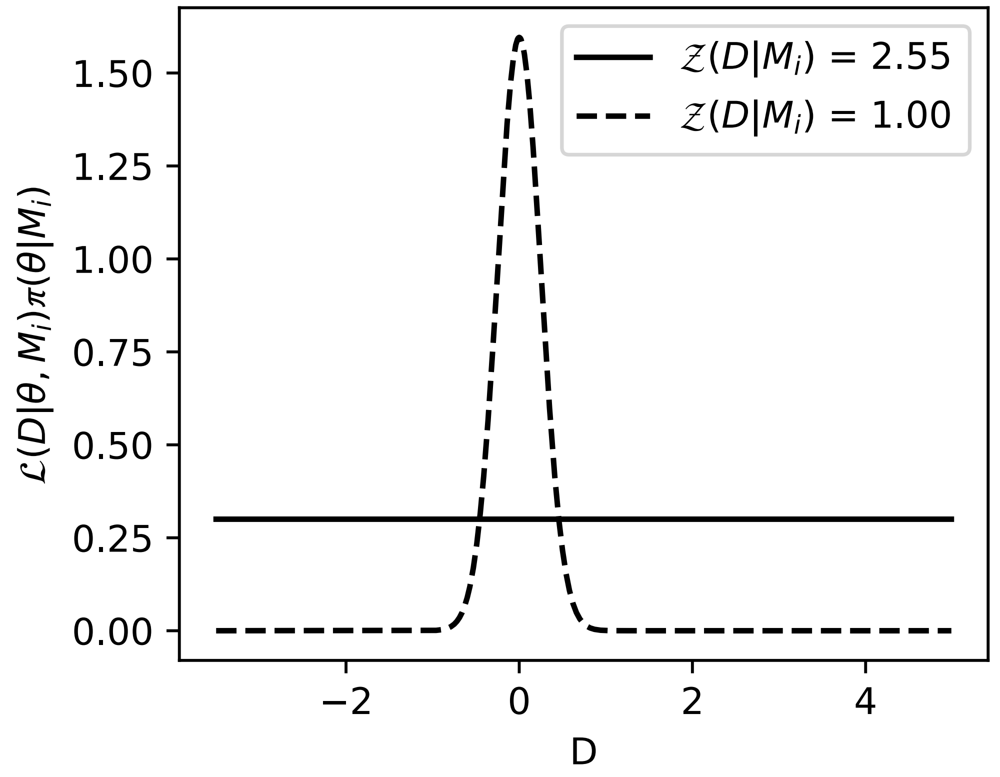
Either used in a suite of models:

$$p(M_i\,|\,D) = \frac{Z(D\,|\,M_i)\pi(M_i)}{\mathscr{F} + \sum_i Z(D\,|\,M_i)\pi(M_i)}$$

$$p(H_0\,|\,D) = \frac{\mathscr{F}}{\mathscr{F} + \sum_i Z(D\,|\,M_i)\pi(M_i)}$$

$$p(M_i\,|\,D, \hat{\theta}_i) = \frac{p(D\,|\,\hat{\theta}_i, M_i)\,p(\hat{\theta}_i\,|\,M_i)\pi(M_i)}{\mathscr{F} + \sum_i p(D\,|\,\hat{\theta}_i, M_i)\,p(\hat{\theta}_i\,|\,M_i)\pi(M_i)}$$

$$p(H_0\,|\,D) = \frac{\mathscr{F}}{\mathscr{F} + \sum_i p(D\,|\,\hat{\theta}_i, M_i)\,p(\hat{\theta}_i\,|\,M_i)\pi(M_i)}$$



Tom J Wilson @onoddil

The Bayesian Posterior Probability

Evidence-based posterior

E.g. Gibson (2014)

$$p(M_i\,|\,D) = \frac{Z(D\,|\,M_i)\,\pi(M_i)}{\sum_i Z(D\,|\,M_i)\,\pi(M_i)}$$

Maximum likelihood-based posterior

$$p(M_i\,|\,D,\hat{\theta}_i) = \frac{p(D\,|\,\hat{\theta}_i, M_i)\,p(\hat{\theta}_i\,|\,M_i)\pi(M_i)}{\sum_i p(D\,|\,\hat{\theta}_i, M_i)\,p(\hat{\theta}_i\,|\,M_i)\pi(M_i)}$$

The "null hypothesis" could be treated as a "fire extinguisher", "to be held in abeyance until needed".

Jaynes (2003), "Probability Theory: The Logic of Science"

Either used in a suite of models:

$$p(M_i\,|\,D) = \frac{Z(D\,|\,M_i)\pi(M_i)}{\mathscr{F} + \sum_i Z(D\,|\,M_i)\pi(M_i)}$$

$$p(H_0\,|\,D) = \frac{\mathscr{F}}{\mathscr{F} + \sum_i Z(D\,|\,M_i)\pi(M_i)}$$

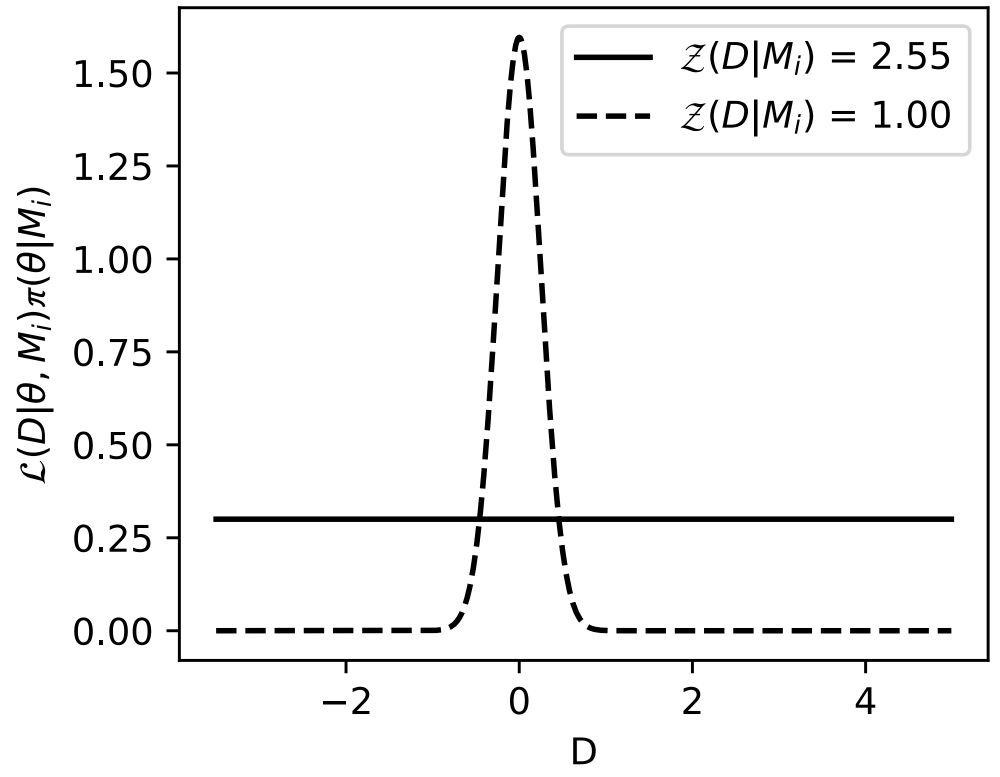
$$p(M_i\,|\,D,\hat{\theta}_i) = \frac{p(D\,|\,\hat{\theta}_i, M_i)\,p(\hat{\theta}_i\,|\,M_i)\pi(M_i)}{\mathscr{F} + \sum_i p(D\,|\,\hat{\theta}_i, M_i)\,p(\hat{\theta}_i\,|\,M_i)\pi(M_i)}$$

$$p(H_0\,|\,D) = \frac{\mathscr{F}}{\mathscr{F} + \sum_i p(D\,|\,\hat{\theta}_i, M_i)\,p(\hat{\theta}_i\,|\,M_i)\pi(M_i)}$$

or in the parameterisation of a single model:

$$Z(D\,|\,M_i) = \int_{\text{all}\,\theta} \mathscr{L}(D\,|\,\theta, M_i)\pi(\theta\,|\,M_i)\,\mathrm{d}\theta + \mathscr{F}'$$

$$p(\theta\,|\,D, M_i) = \frac{\mathscr{L}(D\,|\,\theta, M_i)\pi(\theta\,|\,M_i)}{Z(D\,|\,M_i)}$$

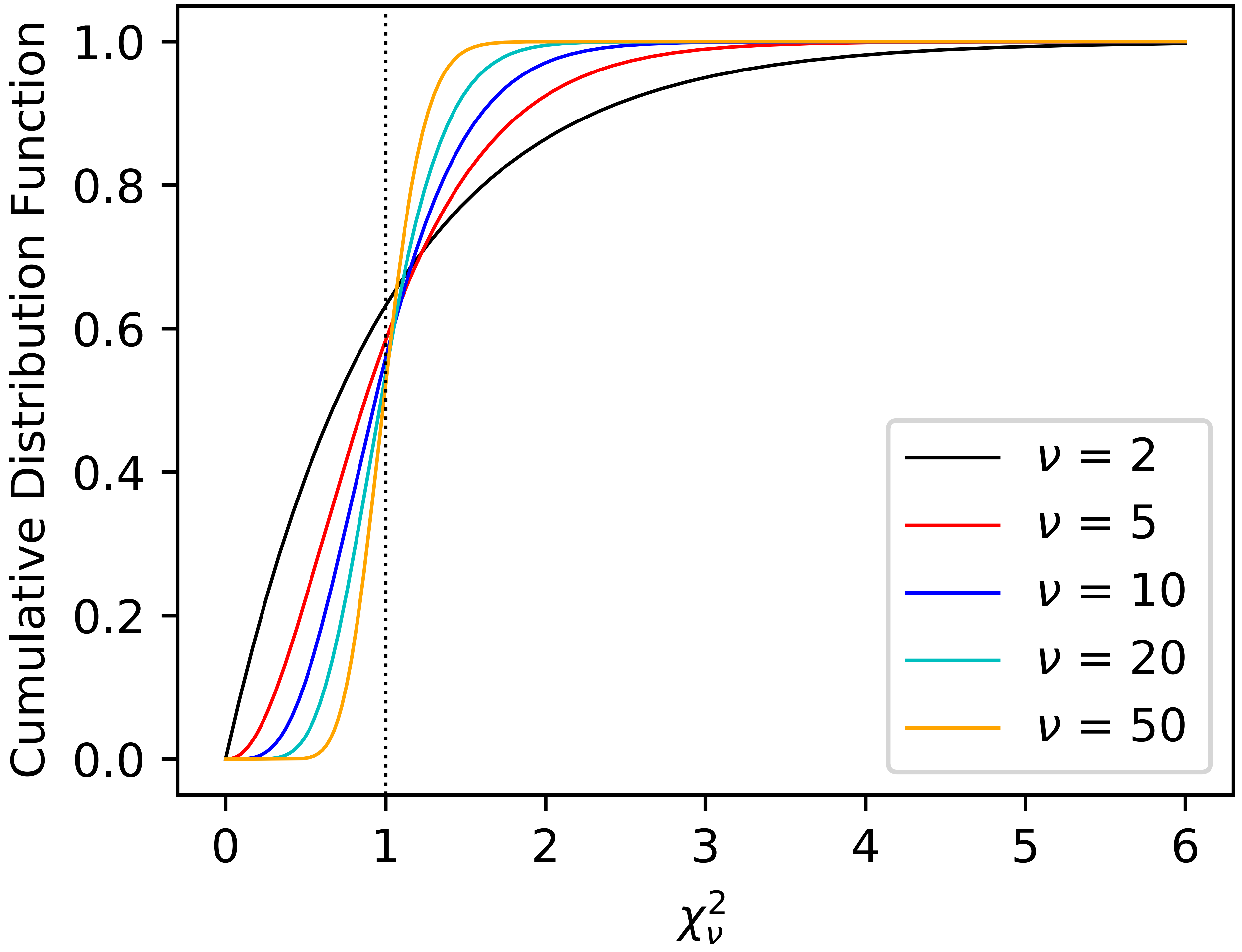


Tom J Wilson @onoddil

The Chi-Squared Statistic

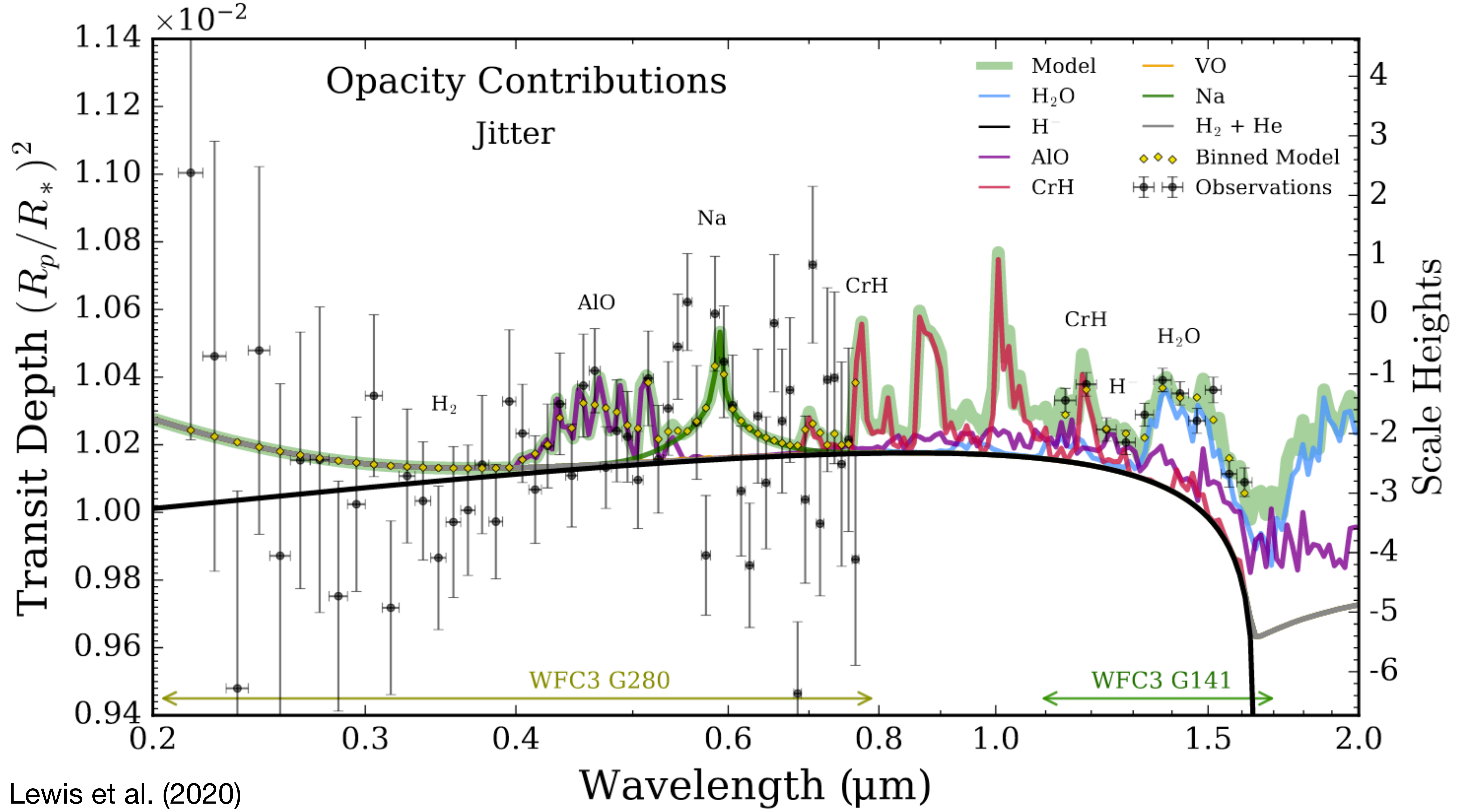
$$Q = \sum_{i}^{n} \frac{(f(x_i) - y_i)^2}{\sigma_i^2} \sim \chi^2(n-k) \qquad P(\chi^2 \leq x) = F(x;k) = \frac{\gamma(\nu/2,\, x/2)}{\Gamma(\nu/2)}$$

The *null hypothesis*, H_0, describes the idea that chance alone is responsible for one's results — in this case, the normalised residuals to the fit.

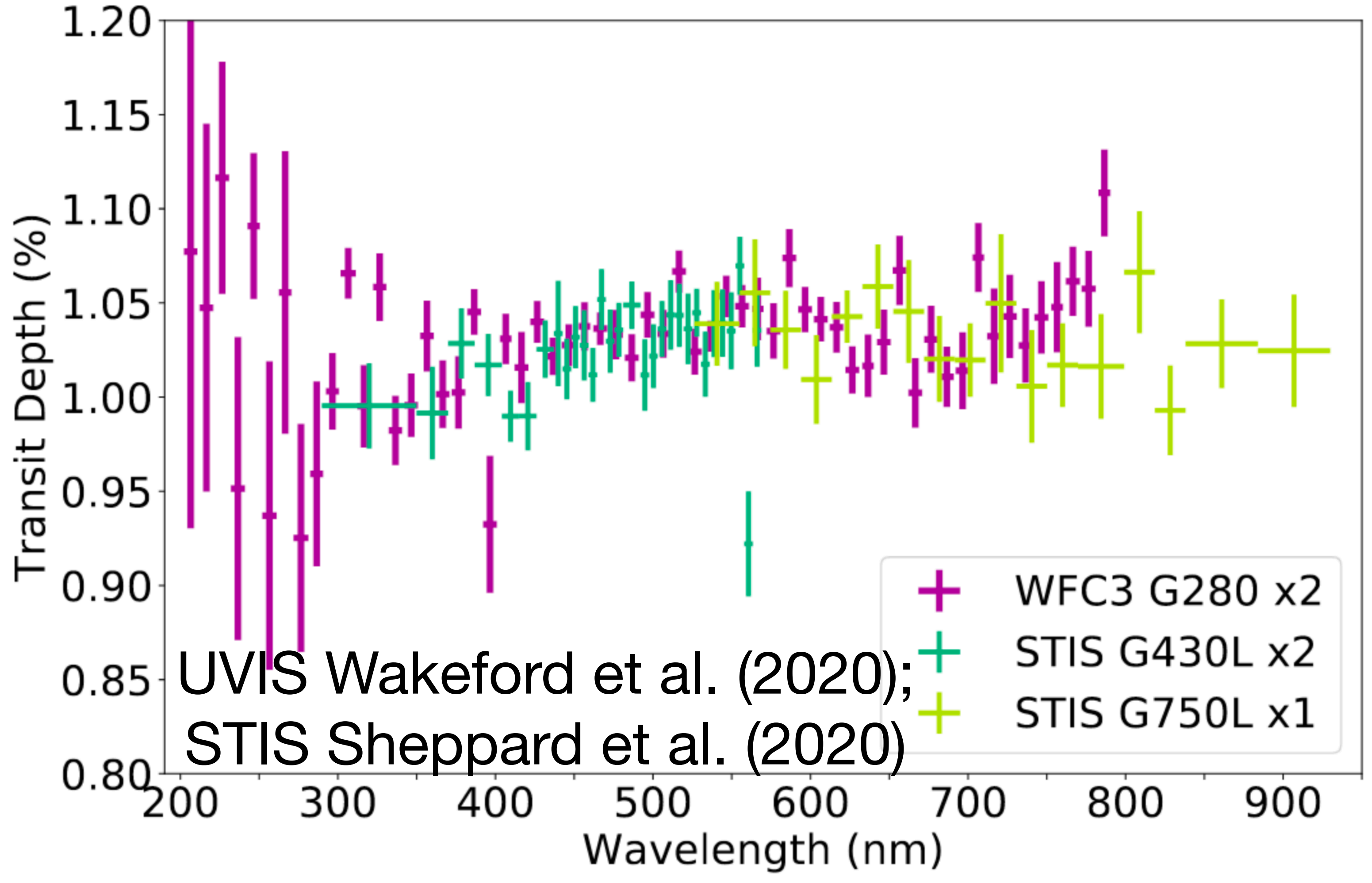


Always quote the (reduced) chi-squared AND degrees of freedom, and consider the rejection of the null hypothesis

WFC3/UVIS and HAT-P-41b



Lewis et al. (2020)

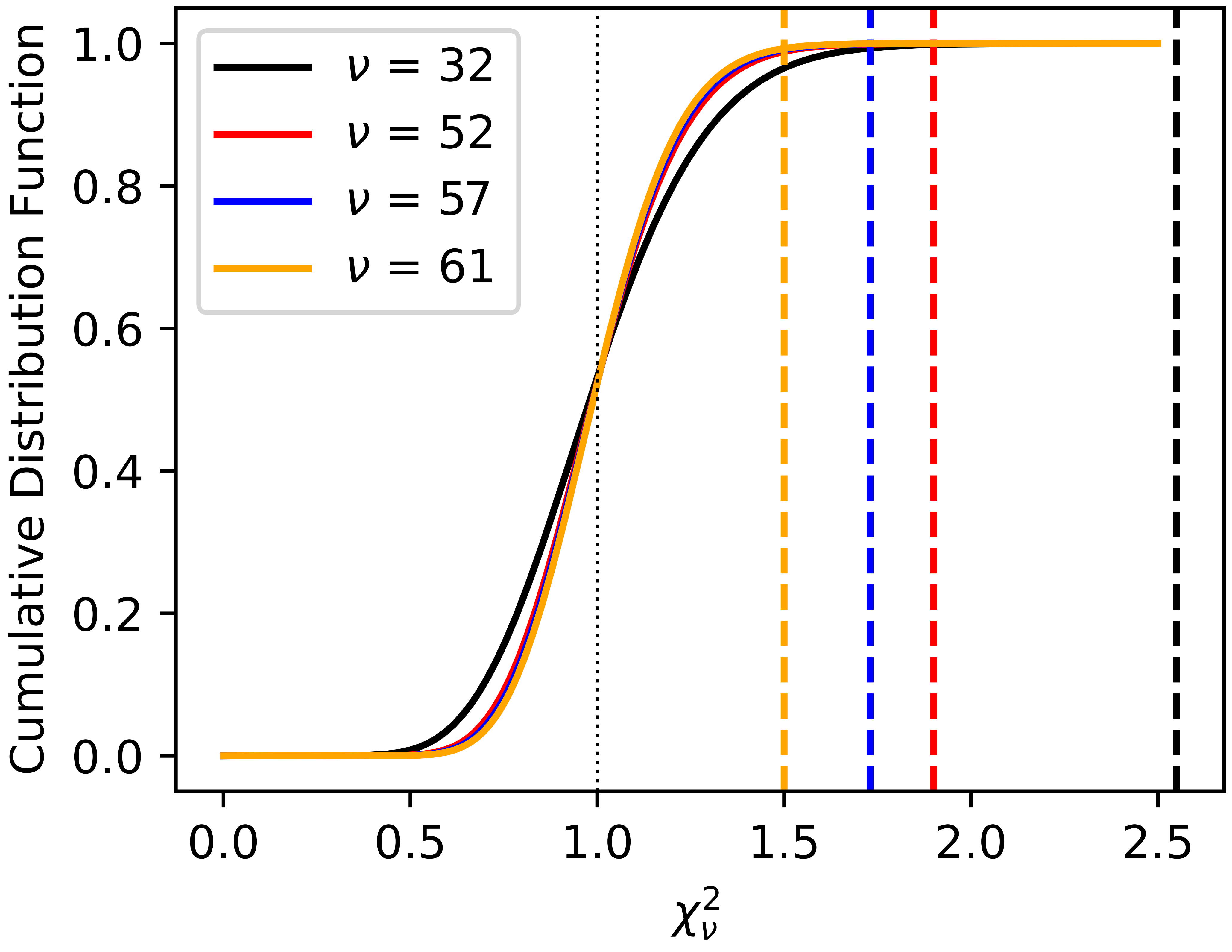


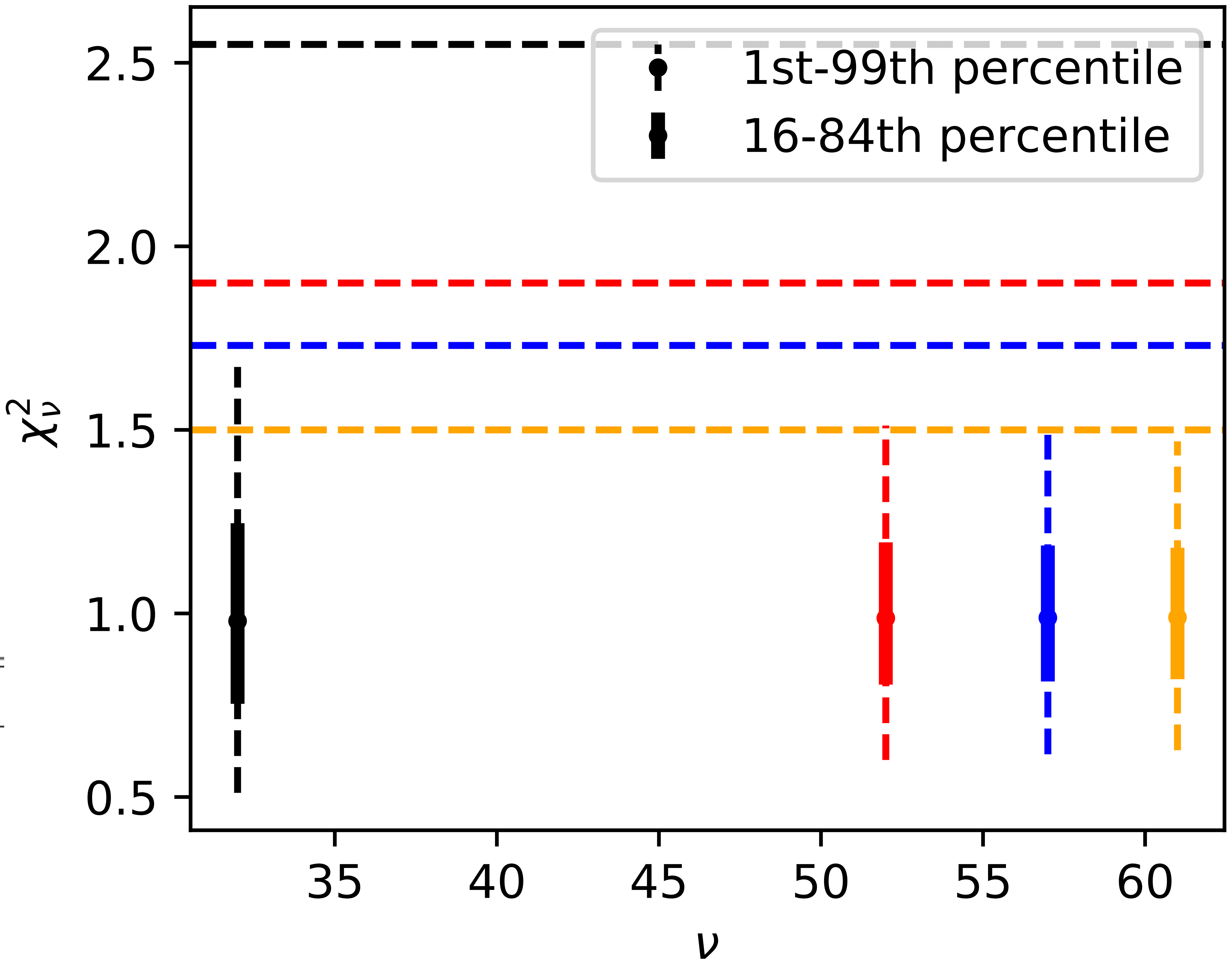
Previous typical DoFs of ~35 —
***JWST* will have hundreds of DoF!**

Data Reduction	Jitter				Marginalization			
Retrieval	POSEIDON	NEMESIS	ATMO	‘Minimal’	POSEIDON	NEMESIS	ATMO	‘Minimal’
d.o.f.	32	52	57	61	32	52	57	61

Tom J Wilson @onoddil

Model-Data Tensions





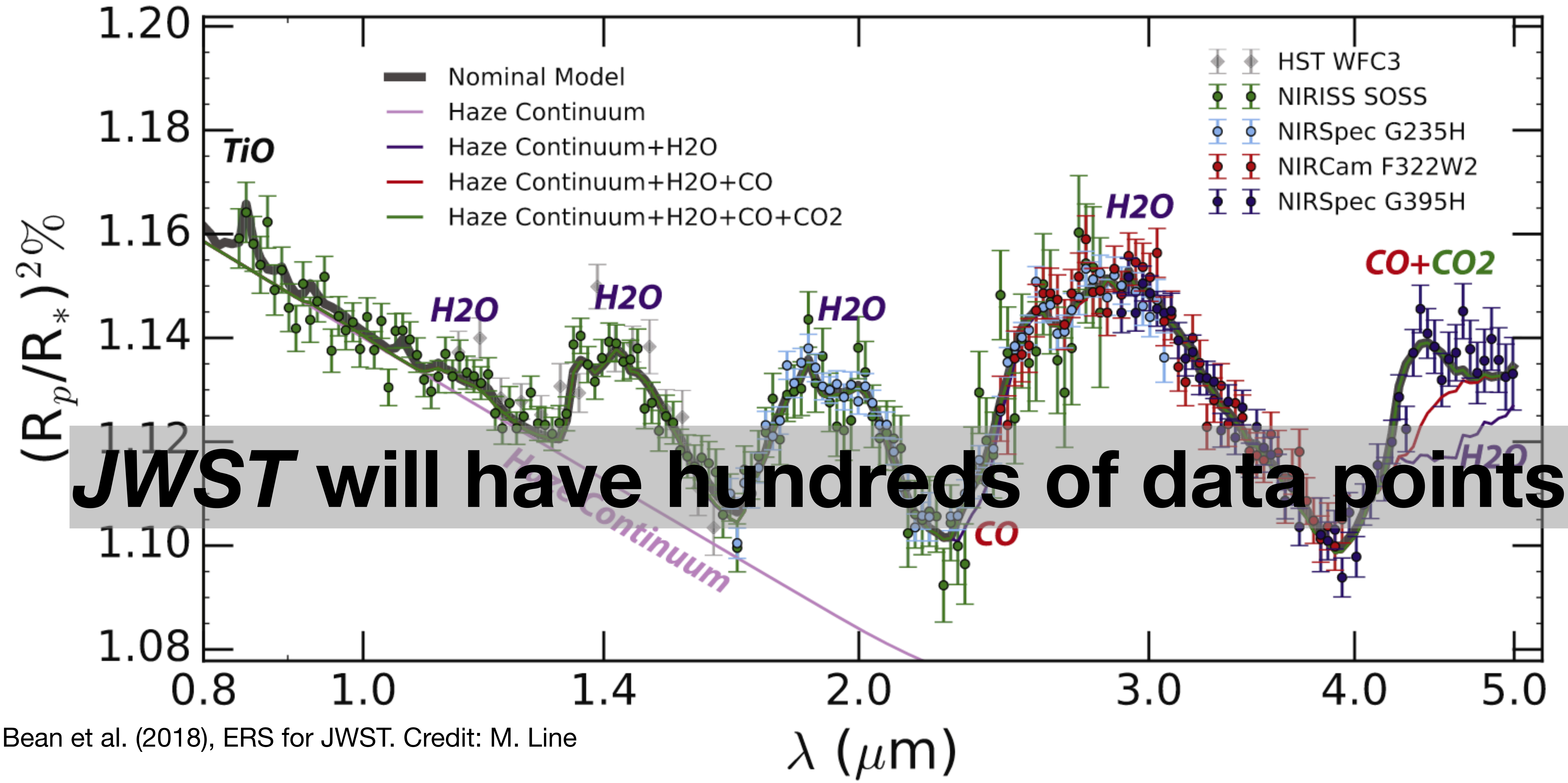
Data Reduction	**Jitter**			
Retrieval	POSEIDON	NEMESIS	ATMO	'Minimal'
Statistics				
ln(Evidence)	473.9	159.8	473.3	478.9
$\chi^2_{\nu,\,\mathrm{min}}$	2.55	1.90	1.73	1.50
N_{param}	37	17	12	8
d.o.f.	32	52	57	61

Lewis et al. (2020)

Reduced chi-squared of "only" 1.5 rejected at >99% probability with increase in dataset size!

Tom J Wilson @onoddil

JWST: An Analysis Turning Point



JWST will have hundreds of data points, not tens

Bean et al. (2018), ERS for JWST. Credit: M. Line

Tom J Wilson @onoddil

Conclusions

What's the best model fit to the data?

(And what do "best" and "model" mean anyway?)

- **"Model" versus "parameterisation" important; we probably care about parameters, not model choice**
  - **Model comparison a shortcut for comparing individual chemical abundances, e.g.**
- **Evidence ratios assume at least one model is correct**
- **Chi-squared CDF or "null hypothesis" can inform on the probability that given model and parameterisation are probable explanations of the dataset, instead of just most likely of choices**
  - **What are the chances that something else is needed to explain these data?**
  - **Differences can be in unexplained data reduction systematics or missing model physics, e.g.**
- **Caution must be given when interpreting Bayesian classifier relative model rankings with increasing precision and numbers of data points — especially for *JWST***
  - **Say "H_2O favoured over its non-inclusion at the 5-sigma level" and "These parameters and model reject the null hypothesis of random chance residuals with 60% probability"**
- **Always quote the chi-squared, the degrees of freedom, and the probability of chi-squared!**