Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **509,** 289–299 (2022)
Advance Access publication 2021 October 14

https://doi.org/10.1093/mnras/stab2941

# Atmospheric characterization of hot Jupiters using hierarchical models of *Spitzer* observations

Dylan Keating [1]★ and Nicolas B. Cowan [1,2]★

[1] *Department of Physics, McGill University, Montréal, QC H3A 2T8, Canada*
[2] *Department of Earth and Planetary Sciences, McGill University, Montréal, QC H3A 2T8, Canada*

Accepted 2021 October 7. Received 2021 October 7; in original form 2021 February 25

## ABSTRACT

The field of exoplanet atmospheric characterization is trending towards comparative studies involving many planetary systems, and using Bayesian hierarchical modelling is a natural next step. Here we demonstrate two use cases. We first use hierarchical modelling to quantify variability in repeated observations by reanalysing a suite of 10 *Spitzer* secondary eclipse observations of the hot Jupiter XO-3 b. We compare three models: one where we fit 10 separate eclipse depths, one where we use a single eclipse depth for all 10 observations, and a hierarchical model. By comparing the widely applicable information criterion of each model, we show that the hierarchical model is preferred over the others. The hierarchical model yields less scatter across the suite of eclipse depths – and higher precision on the individual eclipse depths – than does fitting the observations separately. We find that the hierarchical eclipse depth uncertainty is larger than the uncertainties on the individual eclipse depths, which suggests either slight astrophysical variability or that single eclipse observations underestimate the true eclipse depth uncertainty. Finally, we fit a suite of published dayside brightness measurements for 37 planets using a hierarchical model of brightness temperature versus irradiation temperature. The hierarchical model gives tighter constraints on the individual brightness temperatures than the non-hierarchical model. Although we tested hierarchical modelling on *Spitzer* eclipse data of hot Jupiters, it is applicable to observations of smaller planets like hot Neptunes and super-Earths, as well as for photometric and spectroscopic transit or phase-curve observations.

**Key words:** techniques: photometric – planets and satellites: individual: XO-3 b.

## 1 INTRODUCTION

Although the *Spitzer Space Telescope* was not designed for exoplanet science, it was a workhorse for the field (for a recent review, read Deming & Knutson 2020). In particular, observations of exoplanet transits, secondary eclipses, and phase curves with *Spitzer's* Infrared Array Camera (IRAC) have been used to characterize the atmospheres of over a hundred transiting planets.

Substantial progress has been made towards a statistical understanding of exoplanetary atmospheres (Cowan & Agol 2011; Schwartz & Cowan 2015; Sing et al. 2016; Schwartz et al. 2017; Parmentier & Crossfield 2018; Zhang et al. 2018; Keating, Cowan & Dang 2019; Baxter et al. 2020; Keating et al. 2020; Bell et al. 2021). Many of the planets in these studies had been analysed using disparate reduction and analysis pipelines, but researchers have started uniformly analysing observations of multiple planets using a single pipeline. Garhart et al. (2020) independently reduced and analysed 78 eclipse depths from 36 planets and found that hotter planets had higher brightness temperatures at 4.5 μm than at 3.6 μm. Bell et al. (2021) reanalysed every available *Spitzer* 4.5 μm hot Jupiter phase curve using an open-source reduction and analysis pipeline, confirming several previously reported trends.

In this work, we outline a complementary way to further the statistical understanding of exoplanet atmospheres: fitting measurements from multiple planets simultaneously using hierarchical models to robustly infer trends.

### 1.1 *Spitzer* systematics

Exoplanet observations taken with *Spitzer*'s IRAC (Fazio et al. 2004) are dominated by systematics noise. The systematics are driven by intrapixel sensitivity variations on the detector and by now are well characterized (Ingalls et al. 2016). Detector systematics are typically fitted simultaneously with the astrophysical signal of interest. Each transit, secondary eclipse, and phase curve yield information about the IRAC detector sensitivity, but typically this information is not shared between observations.

Since the *Spitzer* systematics are a function of the centroid location on the pixel, efforts have been made to map the detector sensitivity independently using observations of quiet stars (Ingalls et al. 2012; Krick et al. 2020; May & Stevenson 2020). The flux of a calibration star should be constant as a function of time, so any deviation must be due to the centroid moving across the detector as the telescope pointing drifts. A crucial assumption for this approach is that the

★ E-mail: dylan.keating@mail.mcgill.ca (DK); nicolas.cowan@mcgill.ca (NBC)

*Spitzer* systematics do not vary with time, and that they are not dependent on the brightness of the star. Ingalls et al. (2012) and May & Stevenson (2020) approached the problem by explicitly calculating the detector sensitivity, while Krick et al. (2020) used a machine learning technique called random forests to look for patterns in the systematics.

Other approaches do not assume anything explicit about the detector sensitivity. Independent component analysis (Waldmann 2012; Morello et al. 2014; Morello, Waldmann & Tinetti 2016) separates the signal into additive subcomponents using blind source separation, with the idea being that one of these signals is the astrophysical signal. In another approach, Morvan et al. (2020) used the baseline signal before and after a transit to learn and predict the in transit detector systematics using a machine learning technique known as Long short-term memory networks.

In this work, we opted to parametrize and fit the detector systematics simultaneously with the astrophysical signal to account for any correlations between the two.

### 1.2 Hierarchical models

Bayesian hierarchical models (Gelman et al. 2014) are routinely used in other fields because they offer a natural way to infer higher level trends in a data set and can increase measurement precision. They are gaining traction in exoplanet studies: for example, to study the mass–radius (Teske et al. 2021) and mass–radius–period (Neil & Rogers 2020) relations, and radius inflation of hot Jupiters (Sarkis et al. 2021; Thorngren et al. 2021). Hierarchical models have not yet been applied to atmospheric characterization of exoplanets.

There is one major difference between a typical Bayesian model and a hierarchical one. In a traditional Bayesian model, we estimate the probability distribution of our model parameters given our observed data and the prior probability of each model parameter. The prior distribution encodes our previous knowledge about the most likely values of the parameters and is specified before fitting the model. In a hierarchical model, however, the prior distributions themselves are parametrized using so-called hyperparameters. The hyperparameters become part of the model and are fitted simultaneously with the other parameters of interest. As we explain in the next section, this naturally represents how our intuition pools information across observations. It also helps to tame models by compromising between overfitting and underfitting.

Hierarchical models should be used whenever the data allow us to refine our knowledge of the prior distribution, which happens when a certain quantity is measured multiple times. A natural example in exoplanet science is repeated *Spitzer* observations of the same target. To demonstrate, we start with the archetypical suite of 10 *Spitzer* IRAC channel 2 (4.5 μm) secondary eclipses of the eccentric hot Jupiter XO-3 b (Wong et al. 2014). Below we explain the model and present our results. Afterwards, we show how we extend the model to fit multiple eclipses from different planets simultaneously and present results from fitting the eclipse data from Garhart et al. (2020) with a hierarchical model.

## 2 HIERARCHICAL MODEL OF XO-3 B ECLIPSES

In the *Spitzer* data challenge, several groups analysed 10 secondary eclipses of XO-3 b in order to test the repeatability and accuracy of various decorrelation techniques (Ingalls et al. 2016). The reduced archival data from the data challenge are publicly available, so we downloaded them rather than reducing them ourselves.

For XO-3 b and other planets with repeated secondary eclipse observations, the eclipses have usually been fitted separately from one another, with a separate eclipse depth parameter for each observation (Ingalls et al. 2016; Kilpatrick et al. 2020). In other cases, a single eclipse depth parameter has been used to simultaneously fit multiple secondary eclipse measurements (Wong et al. 2014). This is also what is typically done for phase curves that are bracketed by two eclipses (Cowan et al. 2012; Bell et al. 2021).

However, neither approach quite matches what our intuition tells us. Because we are measuring the same thing each time, fitting the eclipse observations separately amounts to overfitting the individual observations, and fitting a single eclipse parameter amounts to underfitting all of the observations. If we observe one secondary eclipse, we would expect that the next one we observe would have a similar – but not identical – depth, due to measurement uncertainty, if not astrophysical variability. The second eclipse we observe would also change our beliefs about the first one. Each measurement of the planet's eclipse depth can be thought of as a draw from a distribution, with some variance. With enough measurements, the shape of this distribution can be inferred. A hierarchical model naturally takes all of this into account by fitting for the parameters that describe the higher level distribution simultaneously with the astrophysical signal of each observation.

Bayesian analysis requires us to specify priors on the parameters we are trying to infer. We can write down our prior on the $i$th eclipse depth as

$$D_i \sim \mathcal{N}(\mu, \sigma), \tag{1}$$

where we have used the tilde shorthand ($\sim$) to mean that the eclipse depth is drawn from a normal distribution centred on $\mu$ with a standard deviation of $\sigma$; $\mu$ and $\sigma$ are hyperparameters. In a non-hierarchical model, we would specify $\mu$ and $\sigma$ to represent our prior expectations of what $D_i$ could be. After fitting, we would get a separate posterior distribution for each eclipse depth.

In a hierarchical model, we instead make $\mu$ and $\sigma$ parameters and fit them simultaneously with the 10 eclipse depths. We represent our beliefs about hyperparameters $\mu$ and $\sigma$ with hyperpriors. This allows each eclipse observation to inform the others, by pulling the eclipse depths closer to the mode of the distribution of $\mu$. This is known as Bayesian shrinkage. After fitting, we get a posterior distribution for each eclipse depth, as well as for $\mu$ and $\sigma$.

In the limit that $\sigma$ goes to infinity, the hierarchical model is equivalent to the model with completely separate eclipse depths. Likewise when $\sigma$ goes to zero, it is equivalent to the single eclipse depth model. A hierarchical model empirically fits for the amount of pooling based on what is most consistent with the observations.

### 2.1 Priors

Priors are necessary in a fit to encode prior knowledge, as well as to properly sample a model. In all cases, we use weakly informative priors rather than flat, 'uninformative' priors. A flat prior is equivalent to saying that all values of eclipse depth are equally likely, even extremely large, unphysical values. Instead, we chose to place a normal prior with a large standard deviation so that we kept the predicted values within the right order of magnitude. Half-normal priors or wide normal priors are unlikely to introduce much bias into the parameter estimates and can make sampling more efficient. Flat priors are discouraged in practice because we usually have at least some vague knowledge of the range of values a parameter can take (Gelman, Simpson & Betancourt 2017).

### 2.2 Astrophysical model

The astrophysical model for each observation was a secondary eclipse. We used STARRY(Luger et al. 2019) to compute the shape of each eclipse, with the depth and time of eclipse left as free parameters. We fixed the radius of the planet and host star, the orbital period, ratio of semimajor axis to stellar radius, orbital inclination, longitude of periastron, and eccentricity to the literature values.

To get a rough upper limit on the eclipse depth, we used the parametrization of Cowan & Agol (2011) to calculate the maximum dayside temperature, in the limit of a Bond albedo of zero and no heat recirculation:

$$T_{\rm d,max} = T_{\rm eff}\sqrt{\frac{R_\star}{a}}\left(\frac{2}{3}\right)^{1/4}. \tag{2}$$

Here $T_{\rm eff}$ is the stellar effective temperature, and $a/R_\star$ is the ratio of semimajor axis to stellar radius. We note that this equation assumes a circular orbit, while XO-3 b is on an eccentric orbit ($e = 0.28$; Bonomo et al. 2017). None the less, it allows us to get an order of magnitude estimate of the eclipse depth.

The above temperature can be converted to an eclipse depth using

$$D = \frac{B(\lambda, T_{\rm d})}{B(\lambda, T_{\star,\,4.5\,\mu{\rm m}})}\left(\frac{R_{\rm p}}{R_\star}\right)^2, \tag{3}$$

where $B$ is the Planck function, and $T_{\star,\,4.5\,\mu{\rm m}}$ is the brightness temperature of the star at 4.5 μm, which we calculated by integrating PHOENIX models (Allard, Homeier & Freytag 2011) over the *Spitzer* bandpass (Baxter et al. 2020). We represent the eclipse depth when $T_{\rm d} = T_{\rm d,\,max}$ by $D_{\rm max}$.

For the non-hierarchical model, we placed a wide prior on the eclipse depth to prevent biasing the value: $D \sim \mathcal{N}(D_{\rm max}/2, D_{\rm max}/2)$. For the time of eclipse, we let $\tau \sim \mathcal{N}(\Delta t/2, \Delta t/2)$, where $\Delta t$ is the duration of the observation, and time is measured from the start of the observation. We experimented with various priors and found that our resulting fits were consistent and not strongly dependent on the choice of priors.

For the hierarchical model, we used a wide Normal prior for the hierarchical mean: $\mu \sim \mathcal{N}(D_{\rm max}/2, D_{\rm max}/2)$. For the hierarchical standard deviation we used a weakly informative half-Normal prior: $\sigma \sim$ half-$\mathcal{N}(300\,{\rm ppm})$. We then let the individual eclipse depths be drawn from the following higher level distribution: $D_i \sim \mathcal{N}(\mu, \sigma)$.

### 2.3 Detector systematics: Gaussian processes

The IRAC detector sensitivity in channels 1 and 2 depends on the target centroid position on the detector. To parametrize this behaviour, we used a Gaussian process (GP). The advantage of using a GP is that it does not require calculating the detector sensitivity explicitly, in contrast with polynomial models (Cowan et al. 2012) or Bilinearly Interpolated Subpixel Sensitivity (BLISS; Stevenson et al. 2012).

When using GPs, we make the usual assumption that the data are normally distributed, but allow for covariance between data points. The likelihood function can be written as

$$p({\rm data}|\gamma) \sim \mathcal{N}(\mu_{\rm GP}, \Sigma), \tag{4}$$

where $\gamma$ represents the model parameters and independent variables, and $\mu_{\rm GP}$ is the mean function around which the data are distributed. The covariance function, $\Sigma$, is an $n \times n$ matrix, where $n$ is the number of data. The entries along the diagonal of $\Sigma$ are the measurement uncertainties on each datum, which we denote by $\sigma_{\rm phot}$, and the off-diagonal entries are the covariance between data. When the off-diagonal elements are equal to zero, the likelihood function reduces to the usual assumption of independent Gaussian uncertainties.

Although it is computationally intractable to fit each off-diagonal entry of the covariance matrix, they can be parametrized using a kernel function with a handful of parameters. We used the squared exponential kernel employed by Evans et al. (2015):

$$\Sigma_{ij} = A\exp\left[-\left(\frac{x_i - x_j}{l_x}\right)^2 - \left(\frac{y_i - y_j}{l_y}\right)^2\right], \tag{5}$$

where $x$ and $y$ represent the centroid locations on the IRAC detector, in pixel coordinates. The terms $l_x$ and $l_y$ are the covariance length scales, and $A$ is the GP amplitude. The squared exponential kernel has the intuitive property that locations on the detector pixel that are close together should have similar sensitivity. If the length scales are fixed by the user rather than fitted for, this boils down to the Gaussian kernel regression of Ballard et al. (2010), Knutson et al. (2012), and Lewis et al. (2013).

With no loss of generality, we can let the mean function be zero, and instead fit the residuals between the astrophysical model and observations that gives

$$p({\rm residuals}|\gamma) \sim \mathcal{N}(0, \Sigma). \tag{6}$$

We placed weakly informative inverse gamma priors on the length scales. We chose the parameters such that 99 per cent of the prior probability was between length scales 0 and 1, measured in pixels. This gives $p(l_x), p(l_y) \sim {\rm InverseGamma}(\alpha{=}11, \beta = 5)$.

For the amplitude $A$, we used a weakly informative half-Normal prior: $A \sim$ half-$\mathcal{N}(0, \Delta F/3)$, where $\Delta F$ is the range of observed flux values. By using a half-normal prior, we weakly constrain the scale of the GP amplitude without introducing bias. We placed the same prior on the white noise uncertainty $\sigma_{\rm phot}$.

### 2.4 The posterior

For each eclipse, we fit the eclipse depth $D$, time of eclipse $\tau$, photometric uncertainty $\sigma_{\rm phot}$, GP length scales $l_x$ and $l_y$, and GP amplitude $A$. The planet-to-star flux ratio as a function of time $t$ is given by $F$. The $x$ and $y$ centroid locations are included as covariates. We also fit for the hierarchical eclipse depth mean $\mu$, and standard deviation $\sigma$.

The likelihood function for one eclipse is given by

$$p(F|t, x, y, \sigma_{\rm phot}, D, \mu, \sigma, \tau, l_x, l_y), \tag{7}$$

and the prior is

$$p(t, x, y, \sigma_{\rm phot}, D, \mu, \sigma, \tau, l_x, l_y). \tag{8}$$

We form the posterior function by multiplying the likelihood and prior together:

$$p(D, \mu, \sigma, \theta|F) \propto p(F|D, \mu, \sigma, \theta)p(D, \mu, \sigma, \theta), \tag{9}$$

where we have used $\theta = \{t, x, y, \sigma_{\rm phot}, \tau, l_x, l_y\}$ for simplicity.

First, note that the likelihood function does not depend on the hierarchical parameters directly, so we can remove $\mu$ and $\sigma$ from the brackets of the likelihood function. Second, the eclipse depth depends on the hierarchical parameters in this way:

$$p(D|\mu, \sigma) = \frac{p(D, \mu, \sigma)}{p(\mu, \sigma)}, \tag{10}$$

where we have made use of Bayes' theorem.

This means we can rewrite the likelihood and prior to obtain

$$p(D, \mu, \sigma, \theta|F) \propto p(F|D, \theta)p(D|\mu, \sigma)p(\mu, \sigma, \theta). \tag{11}$$

It is this refactoring that makes a hierarchical model different from a non-hierarchical one.

To form the posterior of the full hierarchical model, we multiply the individual eclipse posteriors together:

$$\prod_{i=1}^{n} p(D_i, \mu, \sigma, \theta_i | F_i)$$

$$\propto (p(\mu, \sigma))^n \prod_{i=1}^{n} p(F_i | D_i, \theta_i) p(D_i | \mu, \sigma) p(\theta_i), \tag{12}$$

where $n$ is the number of eclipse observations, 10 in the case of the XO-3 b data set. We have also used the fact that the priors on $\theta_i$ are independent from the priors on the hyperparameters $\mu$ and $\sigma$ to perform the separation $p(\mu, \sigma, \theta_i) = p(\mu, \sigma) p(\theta_i)$.

### 2.5 Hamiltonian Monte Carlo

We used the probabilistic programming package PyMC3 to build and sample from the model. PyMC3 uses Hamiltonian Monte Carlo (HMC), the state-of-the-art Markov chain Monte Carlo (MCMC) algorithm, to perform the sampling. HMC is more efficient than other MCMC algorithms, meaning it can effectively describe the posterior using fewer samples than other MCMC algorithms. For high-dimensional models, and especially for high-dimensional hierarchical models with pathological parameter spaces, the reduction in sample size afforded by HMC is all but necessary (Betancourt & Girolami 2013).

HMC works by treating probabilistic systems as if they are instead physical systems (Betancourt 2017). The chains in an MCMC sampler move through parameter space to estimate the shape of the posterior; an equivalent physical system is the motion of a satellite orbiting a giant planet where the planet represents the mode of the probability distribution. The crucial step in HMC is to transform these trajectories from parameter space to momentum space (i.e. the space of the derivatives of the coordinates) using the Hamiltonian of the system, and sample from that instead by proposing a move in momentum space. By using conservation of energy, the HMC chains tend to remain in regions of high probability. In this way they efficiently traverse the typical set of the distribution, roughly defined as where the most of the probability mass of the posterior is concentrated.

To use HMC, the gradient of the likelihood function, with respect to the parameters, is needed in order to construct the Hamiltonian of the system. PyMC3 does this using THEANO, which is a deep learning library that allows for efficient manipulation of matrices (The Theano Development Team et al. 2016). STARRY is also built on THEANO and allows for analytic expressions and gradients in the general case of eclipse and phase mapping. In our case, because we are dealing with eclipse-only observations and are able to neglect planetary limb darkening, the eclipse expressions reduce to the analytic expressions of Mandel & Agol (2002). Because astrophysical parameters tend to be correlated, we used the dense mass matrix HMC step from the EXOPLANET package (Foreman-Mackey et al. 2021).

The biggest advantage is that HMC can diagnose problematic posteriors or models. Posteriors with pathological regions, such as high curvature, are hard for typical MCMC samplers to explore efficiently; hierarchical models exhibit such pathologies. This can lead to biases in the final results that are hard to diagnose because typical samplers do not have the ability to properly detect and respond to parameter spaces with extreme geometries. When an HMC chain gets stuck in a region of high curvature or otherwise behaves badly, it will diverge to infinity and the sampler keeps track of where this occurred. Divergences can often be eliminated by changing the acceptance probability of the sampler, or by reparametrizing the model.

### 2.6 Model comparison: information criteria

It is common among exoplanet scientists to use the Bayesian information criterion (BIC) or Aikake information criterion (AIC) to perform model comparison and selection (Akaike 1974; Schwarz 1978). Both criteria use the maximum likelihood and a complexity term to penalize overly complex models. These two information criteria describe slightly different things – the AIC measures the relative predictive loss of a set of models, and the BIC measures how close each model is to the true model. In practice (at least in exoplanet science), they typically yield similar conclusions.

A shortcoming shared by BIC and AIC is that they are accurate only when using flat priors, which are not recommended for most models (Gelman et al. 2017). The AIC also assumes that the posterior distributions are multivariate Gaussians. The priors are never flat for hierarchical models, which means we cannot use the AIC or BIC for model comparison.

A more general model comparison tool is the widely applicable information criterion (WAIC; Watanabe 2010). The WAIC is Bayesian, uses the full fit posterior and, critically, makes no assumptions about the shape of the posterior or priors. The WAIC is easily computed from the full fit posterior (McElreath 2020):

$$\mathrm{WAIC}(\mathrm{data}, \Theta) = -2 \left( \mathrm{lppd} - \sum_i \mathrm{var}_\theta \left( \log p \left( F_i \mid D_i, \theta_i \right) \right) \right), \tag{13}$$

where $\Theta$ is the posterior, $F_i$ stands for the $i$th eclipse observation, data refer to the entire suite of observations, and lppd stands for the log-pointwise predictive density,

$$\mathrm{lppd} = \sum_i \log \left( \frac{1}{S} \sum_{s=0}^{S} p \left( F_i \mid D_i, \theta_{i,s} \right) \right), \tag{14}$$

where $S$ is the number of samples and $\theta_{i,s}$ is the $s$th set of parameters for the $i$th observation. The log predictive pointwise density is an estimate of how well the model would fit new, unseen data. The second term in the WAIC expression is a penalty term that penalizes overly complex models.

Once MCMC sampling has finished, the WAIC can be computed in a few lines of code using the MCMC chains. It is also possible to compute the standard error of the WAIC, something that is not possible with the BIC or AIC. If the difference in WAIC between two models is significantly larger than the standard error of the difference, then the model with the smaller WAIC is favoured over the other. If the difference in WAIC is smaller than the standard error of the difference, then the models make equally good predictions and there is no evidence to favour one over the other.

Since all modern secondary eclipse, transit, and phase curve analyses use MCMC to sample and store the posterior draws, the WAIC is a better choice than BIC or AIC for model comparison.

### 2.7 Pooling the GP parameters

We hypothesized that fitting a common set of Gaussian process (GP) amplitude and length scales across the suite of eclipses would yield more precise eclipse depths by sharing information about the detector sensitivity across the observations. Because we used HMC, it was

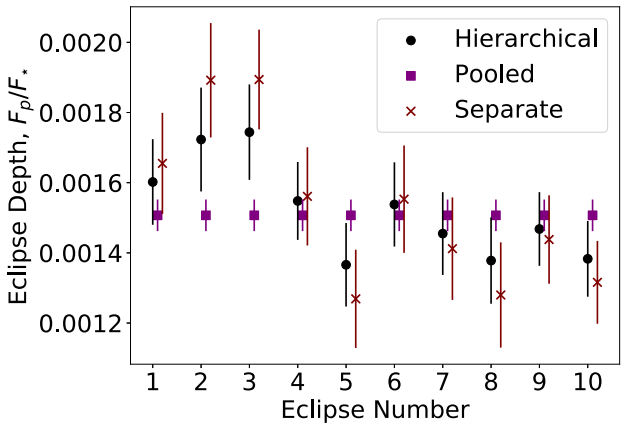


**Figure 1.** Eclipse depths of XO-3 b for the three different models. The pooled model has the lowest eclipse depth uncertainty, but is not the best model according to the WAIC. The hierarchical model is the best model, and yields eclipses that are closer together, with lower uncertainties on the individual eclipse depths than the separate model (15 per cent smaller on average). The hierarchical model represents a compromise between the overfit separate model and the underfit pooled model.

feasible to fully marginalize over the GP hyperparameters. However, we found that the fitted eclipse depths had nearly identical means and standard deviations between the shared and non-shared detector models of XO-3 b. In practice we adopted the shared GP model for XO-3 b because it has fewer parameters and is therefore easier to sample.

### 2.8 Results

We fit the ten 4.5 μm eclipses of XO-3 b with three models: a model where each eclipse observation had its own, separate eclipse depth parameter; one where we used a single, pooled eclipse depth for all the observations; and a hierarchical model. To fit each model, we used 2000 tuning steps to initialize four HMC chains, and obtained 1000 samples for each chain. After sampling, we confirmed that the Gelman–Rubin statistic was close to 1 for all parameter values, and that there were no divergences. The eclipse depths from each model are shown graphically in Fig. 1 and tabulated in Table 1. The best-fitting model for each eclipse observation is shown in Fig. 2. Fig. 3 shows the best-fitting eclipse signal after removing the detector systematics.

We also computed the ΔWAIC values compared to the best-fitting model, shown in Table 2. The hierarchical model had a significantly lower WAIC than the separate model, and a marginally lower WAIC than the single eclipse depth (pooled) model. This suggests that the eclipse depths are indeed different between observations, but are more similar than if we had used a separate eclipse parameter to describe each. This also hints at some variability from observation to observation. Typical hot Jupiters are predicted to show some epoch-to-epoch variability (Komacek & Showman 2020), which is another reason to adopt the hierarchical model over the pooled one.

The mean eclipse depth for the three models is consistent with one another within the uncertainties. In Fig. 1, we see the effects of shrinkage on the eclipse depths. Compared to the separate model, the hierarchical model yields smaller scatter across the suite of eclipse depths and higher precision on the individual eclipse depths. Additionally, the individual uncertainties on the fitted eclipse depths are smaller by 15 per cent on average in the hierarchical fits compared to the separate fits. The individual eclipse depth observations help constrain each other by shrinking the whole suite of eclipse depths towards the grand mean, but not as much as in the completely pooled model.

**Table 1.** Best-fitting eclipse depths from each model of the 10 *Spitzer* IRAC channel 2 eclipses of XO-3 b. The parameters $\mu$ and $\sigma$ are the hierarchical mean and standard deviation for the suite of eclipse depths. With the pooled model, we are implicitly assuming $\sigma = 0$, while for the separate model we are implicitly assuming $\sigma = \infty$. The hierarchical model fits for the amount of pooling, and is preferred over both the completely pooled and separate models.

| Eclipse number | Hierarchical (ppm) | Pooled (ppm) | Separate (ppm) |
|---|---|---|---|
| 1 | $1602 \pm 122$ | $1507 \pm 45$ | $1655 \pm 144$ |
| 2 | $1723 \pm 148$ | $1507 \pm 45$ | $1892 \pm 163$ |
| 3 | $1744 \pm 136$ | $1507 \pm 45$ | $1894 \pm 142$ |
| 4 | $1548 \pm 111$ | $1507 \pm 45$ | $1561 \pm 140$ |
| 5 | $1366 \pm 119$ | $1507 \pm 45$ | $1269 \pm 140$ |
| 6 | $1538 \pm 120$ | $1507 \pm 45$ | $1553 \pm 153$ |
| 7 | $1455 \pm 118$ | $1507 \pm 45$ | $1412 \pm 146$ |
| 8 | $1378 \pm 123$ | $1507 \pm 45$ | $1280 \pm 150$ |
| 9 | $1468 \pm 105$ | $1507 \pm 45$ | $1438 \pm 126$ |
| 10 | $1383 \pm 108$ | $1507 \pm 45$ | $1316 \pm 118$ |
| $\mu$ | $1520 \pm 81$ | $1507 \pm 45$ | $1527 \pm 219$ |
| $\sigma$ | $193 \pm 80$ | 0 | $\infty$ |

## 3 HIERARCHICAL MODEL FOR MULTIPLE PLANETS

While hierarchical modelling is most obviously applicable for repeated measurements of the same planet, we can also extend it to secondary eclipse observations of multiple planets analysed simultaneously. Specifically, we wanted to test the claim from both Garhart et al. (2020) and Baxter et al. (2020) that the 4.5–3.6 μm brightness temperature ratio increases with increasing stellar irradiation for hot Jupiters. We considered the observations from Garhart et al. (2020), the largest data set of uniformly reduced and analysed hot Jupiter secondary eclipses.

We expect that a hot Jupiter's dayside temperature, $T_\mathrm{d}$, is approximately proportional to its irradiation temperature, $T_0 = T_\mathrm{eff}\sqrt{R_\star/a}$. We built a hierarchical model by including this intuition in our hyperprior, and making the hierarchical mean a function of irradiation temperature:

$$\mu_\mathrm{d} = m\,(T_0 - \langle T_0 \rangle) + b,$$

where $\langle T_0 \rangle$ is the average irradiation temperature for the ensemble of planets. In other words, our hierarchical mean is now a line described by a slope and standard deviation in the $T_\mathrm{d}$ versus $T_0$ plane. We represent the scatter about this line using the hyperparameter $\sigma_\mathrm{d}$. The prior on the dayside brightness temperature for a given planet is then $T_{\mathrm{d,p}} \sim \mathcal{N}(\mu_\mathrm{d}, \sigma_\mathrm{d})$.

For this hierarchical model, the hierarchical mean itself depends on two hyperparameters, the slope and intercept of the line, which we fit for simultaneously with the suite of dayside brightness temperatures. We used the following weakly informative priors for the hyperparameters: $m \sim \mathcal{N}(1, 0.5)$, $b \sim \mathcal{N}(2200\,\mathrm{K}, 500\,\mathrm{K})$, and $\sigma_\mathrm{d} \sim \text{half-}\mathcal{N}(500\,\mathrm{K})$.

We included the planets with measurements at both 4.5 and 3.6 μm, which gave a total of 33 planets. Fitting 66 eclipse observations simultaneously with a two-dimensional GP is computationally intractable using the hardware we had available, so we took the

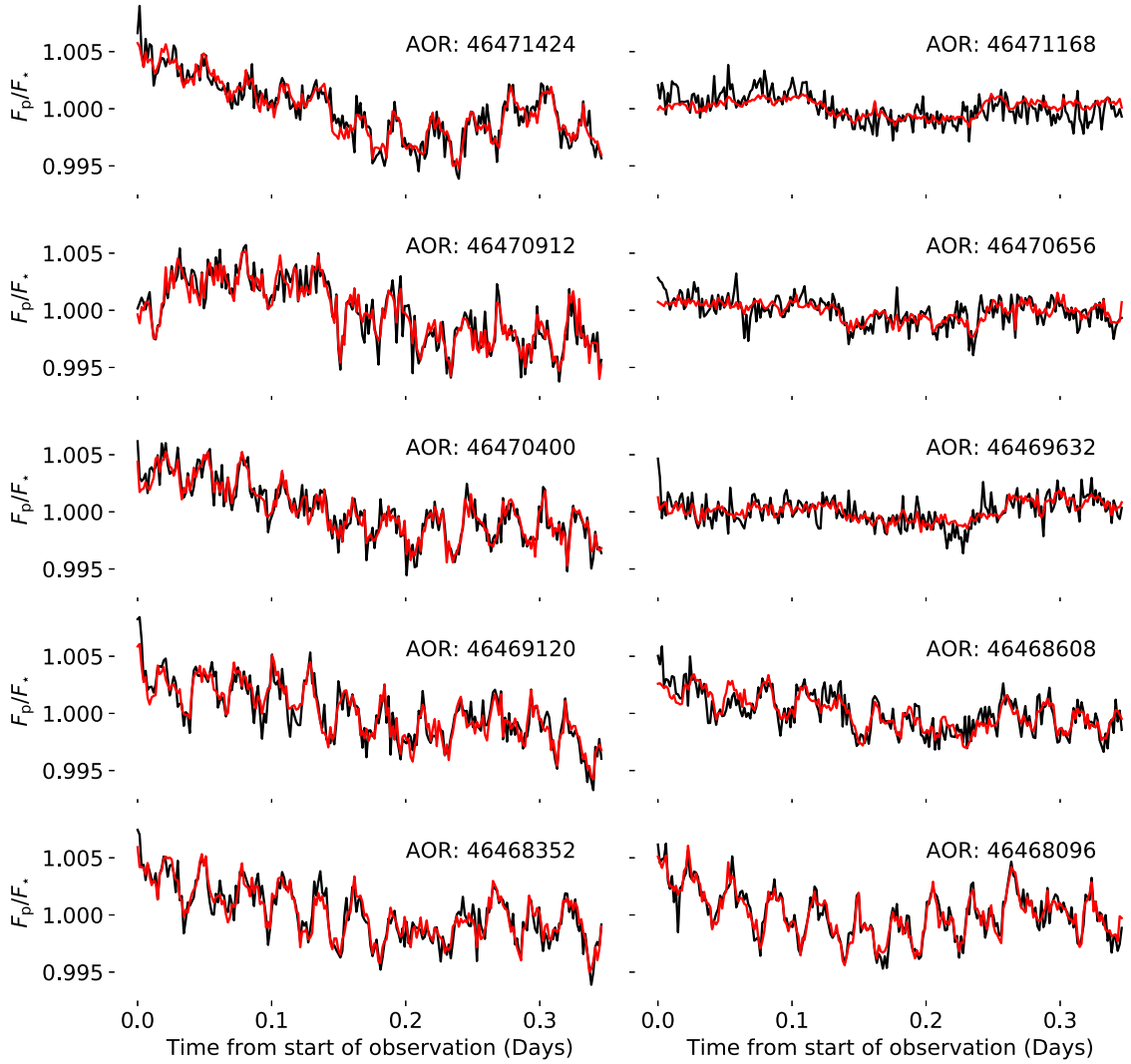


**Figure 2.** The full fit to each of the 10 XO-3 b eclipses, using the hierarchical model for the eclipse depths and a pooled Gaussian process (GP) for the detector systematics.

published measurements at face value rather than refit them. The reported eclipse depths and uncertainties are correct but were not properly propagated when converting to brightness temperature uncertainties (Deming, private communication), so we kept the sample of eclipses and eclipse depths measured by Garhart et al. (2020) but used the brightness temperatures and uncertainties calculated by Baxter et al. (2020). Comparing the two data sets, Garhart et al. (2020) overestimated the brightness temperature uncertainties by about a factor of 2 for each planet.

We again used PYMC3, and first fit a non-hierarchical model as our baseline, using separate $T_{\mathrm{d,\,p}}$ parameters for each eclipse. As expected, this model just reproduces the published dayside temperatures and uncertainties. This also acts as a confidence check that our priors are not biasing the fitted parameters.

Since there are measurements at two different wavelengths, we fit two different versions of the hierarchical model. In the wavelength-dependent model, we allowed the dayside brightness temperature distributions to be different between the two wavelengths, fitting one set of hierarchical parameters for the 4.5 μm measurements, and another set for the 3.6 μm measurements. In the wavelength-independent model, we used a common distribution for all the measurements, and thus one set of hierarchical parameters. We tabulate the refit brightness temperatures in Table 3 and plot them in Fig. 4. We show the refit brightness temperatures scaled by each planets irradiation temperature in Fig. 5.

We show the difference in WAIC values for the three models in Table 4. The wavelength-independent model did slightly better than the wavelength-dependent model ($\Delta\mathrm{WAIC} = 0.54$), however the uncertainty on that difference is 1.35, meaning the models make equally good predictions. In the wavelength-independent model, the dayside brightness temperatures for both channels follow the same slope, or equivalently, the ratio of the slopes for each channel is equal to one. This means we are not detecting – nor ruling out – the trend of increasing brightness temperature ratio versus stellar irradiation reported by Garhart et al. (2020) and Baxter et al. (2020). The best-fitting hierarchical parameters from our wavelength-independent model are $\mu_{\mathrm{d}} = 1.24 \pm 0.06$, $b = 2003 \pm 24$ K, and $\sigma_{\mathrm{m}} = 181 \pm 20$ K.

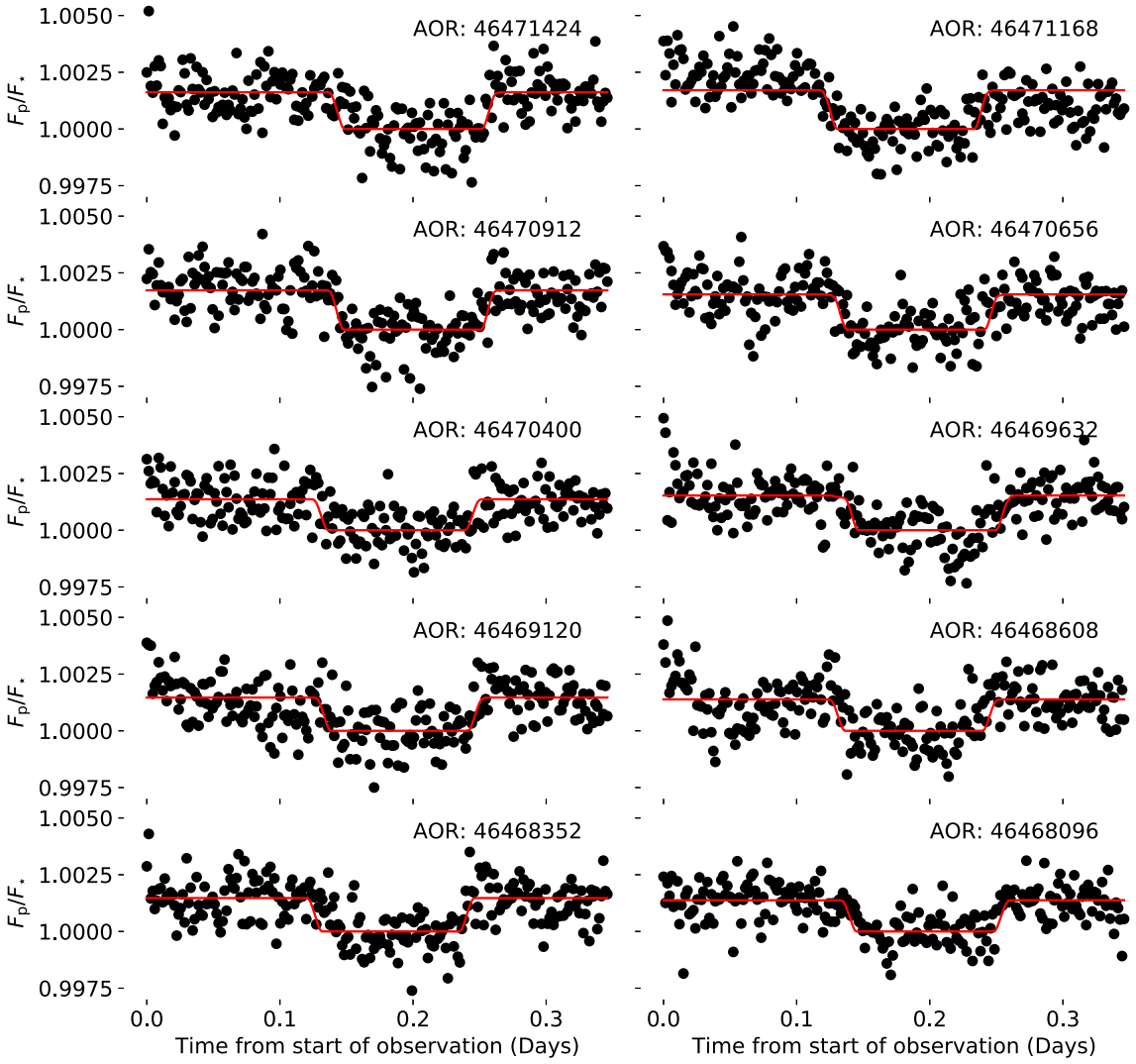


**Figure 3.** The 10 XO-3 b eclipses with the detector systematics removed using a Gaussian process (GP) as a function of stellar centroid location. The eclipse signal is clearly visible in the corrected raw data (represented by the black dots), and the best-fitting eclipse signal for each is shown as the red line.

**Table 2.** Comparison of different information criteria for the three XO-3 b eclipse depth models – lower values are better. For the WAIC we also tabulate the standard error in the difference between each model. Using the BIC and AIC the pooled model is preferred. However, our model does not satisfy the assumptions necessitated by the BIC and AIC. The WAIC is more robust. According to the WAIC, the hierarchical model is the preferred model.

| Model | ΔWAIC | $\sigma_{\Delta\text{WAIC}}$ | ΔAIC | ΔBIC |
|---|---|---|---|---|
| *Hierarchical* | 0.0 | 0.0 | 7.46 | 10.80 |
| Pooled | 8.54 | 5.04 | 0 | 0 |
| Separate | 65.66 | 25.60 | 20.31 | 23.03 |

The difference in WAIC suggests that the non-hierarchical model makes equally good predictions compared to the wavelength-independent hierarchical model (Table 4), which is at odds with expectations that irradiation temperature should determine planetary dayside temperatures.

## 4 DISCUSSION AND CONCLUSIONS

### 4.1 Repeat observations of a single planet

In our reanalysis of the 10 secondary eclipses of XO-3 b, we found that the hierarchical model was favoured over the two non-hierarchical models. This means that the measured eclipse depths are indeed different from epoch to epoch, yet clustered. The biggest difference compared to previous analyses is that we were able to empirically fit for the amount of epoch-to-epoch scatter favoured by the data, and doing this improves the precision on our measurements by 15 per cent on average, because of Bayesian shrinkage. Notably, we found that the hierarchical eclipse depth had a larger standard deviation than the individual measurements, suggesting that measuring just one eclipse depth could lead one to underestimate the true uncertainty compared to the hierarchical approach.

Hierarchical models could improve measurements of other hot Jupiters and other types of planets. Hierarchical models could be used to robustly test the reported variability in the secondary eclipses of the super-Earth 55 Cancri e (Demory et al. 2016; Tamburo et al.

**Table 3.** Results from fitting a wavelength-independent hierarchical model to the Garhart et al. (2020) eclipse data set. The columns $T_{\rm d,\,ch2}$ and $T_{\rm d,\,ch1}$ list the refit dayside brightness temperatures at 4.5 and 3.6 μm.

| Planet | $T_0$ (K) | $T_{\rm d,\,ch2}$ (K) | $T_{\rm d,\,ch1}$ (K) |
|---|---|---|---|
| HAT-P-13 b | 2331 ± 75 | 1739 ± 81 | 1776 ± 79 |
| HAT-P-30 b | 2315 ± 61 | 1763 ± 61 | 1860 ± 49 |
| HAT-P-33 b | 2517 ± 48 | 1912 ± 85 | 1993 ± 63 |
| HAT-P-40 b | 2496 ± 93 | 1867 ± 100 | 1975 ± 113 |
| HAT-P-41 b | 2739 ± 62 | 2171 ± 77 | 2158 ± 124 |
| KELT-2 A b | 2418 ± 44 | 1693 ± 49 | 1861 ± 42 |
| KELT-3 b | 2577 ± 62 | 2006 ± 58 | 2270 ± 57 |
| Qatar-1 b | 1964 ± 61 | 1466 ± 93 | 1409 ± 117 |
| WASP-100 b | 3111 ± 242 | 2362 ± 80 | 2257 ± 74 |
| WASP-101 b | 2198 ± 57 | 1509 ± 56 | 1678 ± 58 |
| WASP-103 b | 3543 ± 110 | 3299 ± 51 | 3005 ± 119 |
| WASP-104 b | 2144 ± 61 | 1779 ± 88 | 1717 ± 70 |
| WASP-12 b | 3654 ± 129 | 2665 ± 42 | 2876 ± 40 |
| WASP-121 b | 3336 ± 86 | 2594 ± 34 | 2370 ± 35 |
| WASP-131 b | 2035 ± 51 | 1174 ± 86 | 1408 ± 98 |
| WASP-14 b | 2636 ± 85 | 2186 ± 83 | 2239 ± 38 |
| WASP-18 b | 3391 ± 103 | 3102 ± 92 | 2917 ± 96 |
| WASP-19 b | 2922 ± 65 | 2273 ± 59 | 2323 ± 53 |
| WASP-36 b | 2403 ± 64 | 1647 ± 125 | 1672 ± 154 |
| WASP-43 b | 1945 ± 112 | 1496 ± 24 | 1660 ± 24 |
| WASP-46 b | 2345 ± 78 | 1910 ± 105 | 1648 ± 146 |
| WASP-62 b | 2018 ± 49 | 1561 ± 58 | 1852 ± 68 |
| WASP-63 b | 2165 ± 64 | 1437 ± 104 | 1586 ± 85 |
| WASP-64 b | 2390 ± 74 | 1705 ± 122 | 2051 ± 79 |
| WASP-65 b | 2100 ± 83 | 1367 ± 131 | 1727 ± 94 |
| WASP-74 b | 2720 ± 75 | 2108 ± 49 | 2003 ± 38 |
| WASP-76 b | 3087 ± 66 | 2471 ± 32 | 2412 ± 28 |
| WASP-77 A b | 2363 ± 44 | 1635 ± 36 | 1689 ± 31 |
| WASP-78 b | 3246 ± 124 | 2579 ± 148 | 2699 ± 123 |
| WASP-79 b | 2492 ± 75 | 1885 ± 52 | 1895 ± 47 |
| WASP-87 b | 3268 ± 96 | 2815 ± 79 | 2673 ± 76 |
| WASP-94 A b | 2127 ± 109 | 1412 ± 49 | 1530 ± 35 |
| WASP-97 b | 2178 ± 59 | 1593 ± 43 | 1723 ± 39 |

**Table 4.** WAIC scores for the three models used to fit the suite of eclipses from Garhart et al. (2020) – lower values are better. We also tabulate the standard error of the difference in WAIC between each model. The three models make equally good predictions according to the WAIC scores.

| Model | ΔWAIC | $\sigma_{\Delta\rm WAIC}$ |
|---|---|---|
| Separate | 0.0 | 0.0 |
| Wavelength independent | 1.13 | 2.54 |
| Wavelength dependent | 1.67 | 2.42 |

2018), or to fit the 12 *Spitzer* eclipses of the recently discovered hot Saturn LTT 9779 b (Dragomir et al. 2020). The published variability constraints for HD 189733 b (Agol et al. 2010) and HD 209458 b (Kilpatrick et al. 2020) could also be revisited with hierarchical models.

Repeated phase-curve observations could benefit from using hierarchical models. The hot Jupiter WASP-43 b has one published (Stevenson et al. 2017; Mendonça et al. 2018; Morello et al. 2019; May & Stevenson 2020; Bell et al. 2021), and two unpublished *Spitzer* phase curves at 4.5 μm. A hierarchical model could be used to better constrain the phase amplitudes and offsets of the three 4.5 μm phase curves by fitting them simultaneously.

To test whether we are seeing the effects of variability or detector systematics, the best approach is to compare planets with repeated observations in both *Spitzer* channels. If certain types of planets have larger hierarchical eclipse standard deviations, the culprit could be time variability that is only exhibited by certain planets. Otherwise, if one *Spitzer* channel tends to show more variability regardless of planet, it suggests that detector systematics are at play. Spectroscopic observations will also be able to break the degeneracy

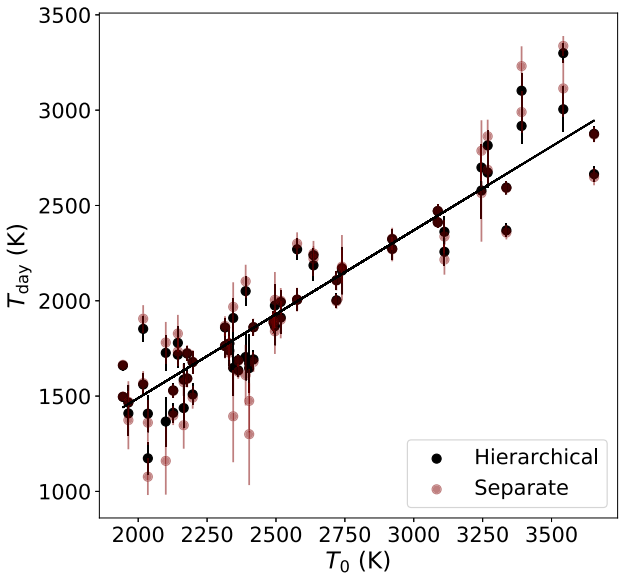


**Figure 4.** Dayside brightness temperatures, at 3.6 and 4.5 μm, for the planets analysed by Garhart et al. (2020), refit with a wavelength-independent hierarchical model. The red dots are the published values and the black line is the best-fitting trend line. The effects of Bayesian shrinkage are evident: the measurements are clustered closer to the line, and the uncertainties on the measurements are reduced.

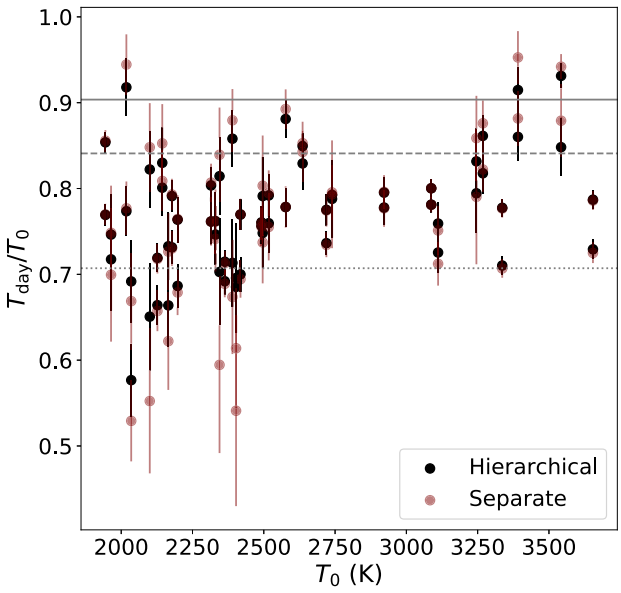


**Figure 5.** Dayside brightness temperatures from Fig. 4, scaled by irradiation temperature. The horizontal lines represent some theoretical limits on dayside temperature assuming a zero Bond albedo: the solid line assumes zero heat recirculation, the dashed line assumes a uniform dayside hemisphere but a temperature of zero on the nightside, and the dotted line assumes a uniform temperature at every location on the planet.

between variability and detector systematics, as would simultaneous measurements with multiple instruments.

### 4.2 Parallel analysis of multiple planets

We showed that our hierarchical model of measurements from multiple planets yields smaller uncertainties on the individual eclipse depths, and tends to shrink the eclipse depths toward the trend line. We did not detect the trend of increasing brightness temperature ratio with increasing stellar irradiation reported by Garhart et al. (2020) and Baxter et al. (2020), nor did we rule it out. The hierarchical models made predictions that were as good as the non-hierarchical model, when comparing the WAIC values.

One possible explanation is that the uncertainties on the eclipse depths are underestimated due to detector systematics or astrophysical variability. This was first suggested by Hansen, Schwartz & Cowan (2014), who concluded that the first generation of *Spitzer* eclipse uncertainties may be underestimated by up to a factor of 3, probably due to inadequate treatment of detector systematics. Hot Jupiter infrared eclipse depths are generally assumed to be the same from epoch to epoch because most general circulation models produce stable circulation patterns (Komacek & Showman 2020), but recent work using high-resolution general circulation models (GCMs) predicts multiple equilibria in hot Jupiter atmospheres and transient planetary-scale storms (Cho, Skinner & Thrastarson 2021). The consequence of such variability, much like detector systematics, is that measuring just a single eclipse for a planet in a given bandpass would lead one to underestimate the uncertainty. Indeed, we found that for XO-3 b, the hierarchical standard deviation was larger than the individual uncertainties by about a factor of 1.5–2. This suggests that if we had observed only one eclipse of XO-3 b, we would have underestimated the eclipse depth uncertainty compared to the estimate from the hierarchical model.

In the context of a hierarchical model, small measurement uncertainties leave less leeway for Bayesian shrinkage. Indeed, repeating our analysis using the larger, albeit miscalculated, uncertainties from Garhart et al. (2020) showed a marked improvement when using the hierarchical model compared to the completely separate model (see Appendix A for the results of that analysis).

Another explanation for the marginal performance of hierarchical models on the Garhart et al. (2020) ensemble of planets is that irradiation temperature is not the sole determinant of planetary dayside temperatures. It is becoming clear that secondary parameters like planetary mass, radius, and rotation rate play important roles in determining atmospheric circulation on hot Jupiters (Keating et al. 2019; Bell et al. 2021). Differences in these parameters could contribute additional planet-to-planet scatter.

In this work, we used the largest subset of *Spitzer* secondary eclipses that had been uniformly reduced and analysed. One obvious extension of our work is to refit the detector systematics and astrophysical signals for all *Spitzer* secondary eclipses using a uniform pipeline. We recommend using a hierarchical model for the dayside brightness temperatures and placing a second level of hierarchy on the planets with repeated eclipses. This would take a prohibitively long time using a two-dimensional GP and conventional hardware like we did for XO-3 b, but it could potentially be done using high-performance or GPU computing. Alternatively, such a fit could be done using an easier-to-compute detector model like Pixel-Level Decorrelation (Deming et al. 2015; Garhart et al. 2020), especially with PyMC3.

In this work, we have shown that hierarchical models are useful when analysing repeated measurements from a single target, or when doing comparative exoplanetology of many targets. Next generation telescopes like *James Webb Space Telescope* (*JWST*) and *Atmospheric Remote-sensing Infrared Exoplanet Large-survey* (*ARIEL*) will make repeated measurements of certain targets, and will both carry out photometric and spectroscopic transit, eclipse, and phase curve surveys for a variety of targets (Bean et al. 2018; Tinetti et al. 2018; Charnay et al. 2021). This will allow for atmospheric characterization of potentially thousands of more exoplanets, from Earth-like planets to ultrahot Jupiters, and we recommend that these comparative surveys incorporate hierarchical modelling to make measurements and predictions that are as robust as possible.

## ACKNOWLEDGEMENTS

This project was conceived at the 'Multi-dimensional characterization of distant worlds: spectral retrieval and spatial mapping' workshop hosted by the Michigan Institute for Research in Astrophysics and spearheaded by Emily Rauscher. We are particularly grateful to David van Dyk for a pedagogical introduction to Bayesian shrinkage. This work is based on observations made with the Spitzer Space Telescope, which is operated by the Jet Propulsion Laboratory, California Institute of Technology under a contract with NASA. We acknowledge support from the McGill Space Institute and l'Institut de recherche sur les exoplanètes. We also acknowledge that McGill University is located on unceded indigenous lands. Tiohtià:ke / Montreal is historically known as a meeting place for First Peoples. Lastly, we have made use of open-source software provided by the PYTHON, ASTROPY, SCIPY, MATPLOTLIB, and PYMC3 communities.

## DATA AVAILABILITY

The reduced photometry for the 10 archival secondary eclipse observations of XO-3 b is freely available at https://irachpp.spitzer.caltech.edu/page/data-challenge-2015. The code used in the XO-3 b reanalysis, and a Jupyter Notebook showing the reanalysis of the Garhart et al. (2020) eclipses can both be found at https://github.com/dylanskeating/HARMONiE.

## REFERENCES

Agol E., Cowan N. B., Knutson H. A., Deming D., Steffen J. H., Henry G. W., Charbonneau D., 2010, ApJ, 721, 1861
Akaike H., 1974, IEEE Trans. Automatic Control, 19, 716
Allard F., Homeier D., Freytag B., 2011, in Johns-Krull C., Browning M. K., West A. A., eds, ASP Conf. Ser. Vol. 448, 16th Cambridge Workshop on Cool Stars, Stellar Systems, and the Sun. Astron. Soc. Pac., San Francisco, p. 91
Ballard S. et al., 2010, PASP, 122, 1341
Baxter C. et al., 2020, A&A, 639, A36
Bean J. L. et al., 2018, PASP, 130, 114402
Bell T. J. et al., 2021, MNRAS, 504, 3316
Betancourt M., 2017, preprint (arXiv:1706.01520)
Betancourt M. J., Girolami M., 2013, preprint (arXiv:1312.0906)
Bonomo A. S. et al., 2017, A&A, 602, A107
Charnay B. et al., 2021, Exp. Astron., in press (arXiv:2102.06523)
Cho J. Y. K., Skinner J. W., Thrastarson H. T., 2021, ApJ, 913, L32
Cowan N. B., Agol E., 2011, ApJ, 729, 54
Cowan N. B., Machalek P., Croll B., Shekhtman L. M., Burrows A., Deming D., Greene T., Hora J. L., 2012, ApJ, 747, 82
Deming D., Knutson H. A., 2020, Nat. Astron., 4, 453
Deming D. et al., 2015, ApJ, 805, 132
Demory B.-O., Gillon M., Madhusudhan N., Queloz D., 2016, MNRAS, 455, 2018
Dragomir D. et al., 2020, ApJ, 903, L6

Evans T. M., Aigrain S., Gibson N., Barstow J. K., Amundsen D. S., Tremblin P., Mourier P., 2015, MNRAS, 451, 680
Fazio G. G. et al., 2004, ApJS, 154, 10
Foreman-Mackey D. et al., 2021, exoplanet-dev/exoplanet: exoplanet v0.5.0. Zenodo, https://doi.org/10.5281/zenodo.4737444
Garhart E. et al., 2020, AJ, 159, 137
Gelman A., Carlin J. B., Stern H. S. B. D. D., Vehtari A., Rubin. D. B., 2014, Bayesian Data Analysis, 3rd edn. CRC Press, Boca Raton, FL
Gelman A., Simpson D., Betancourt M., 2017, Entropy, 19, 555
Hansen C. J., Schwartz J. C., Cowan N. B., 2014, MNRAS, 444, 3632
Ingalls J. G., Krick J. E., Carey S. J., Laine S., Surace J. A., Glaccum W. J., Grillmair C. C., Lowrance P. J., 2012, in Clampin M. C., Fazio G. G., MacEwen H. A., Oschmann J. M., Jr, eds, Proc. SPIE Vol. 8442, Space Telescopes and Instrumentation 2012: Optical, Infrared, and Millimeter Wave. SPIE, Bellingham, p. 84421Y
Ingalls J. G. et al., 2016, AJ, 152, 44
Keating D., Cowan N. B., Dang L., 2019, Nat. Astron., 3, 1092
Keating D. et al., 2020, AJ, 159, 225
Kilpatrick B. M. et al., 2020, AJ, 159, 51
Knutson H. A. et al., 2012, ApJ, 754, 22
Komacek T. D., Showman A. P., 2020, ApJ, 888, 2
Krick J. E., Fraine J., Ingalls J., Deger S., 2020, AJ, 160, 99
Lewis N. K. et al., 2013, ApJ, 766, 95
Luger R., Agol E., Foreman-Mackey D., Fleming D. P., Lustig-Yaeger J., Deitrick R., 2019, AJ, 157, 64
McElreath R., 2020, Statistical Rethinking: A Bayesian Course with Examples in R and Stan, 2nd edn. CRC Press, Boca Raton, FL
Mandel K., Agol E., 2002, ApJ, 580, L171
May E. M., Stevenson K. B., 2020, AJ, 160, 140
Mendonça J. M., Malik M., Demory B.-O., Heng K., 2018, AJ, 155, 150
Morello G., Waldmann I. P., Tinetti G., Peres G., Micela G., Howarth I. D., 2014, ApJ, 786, 22
Morello G., Waldmann I. P., Tinetti G., 2016, ApJ, 820, 86
Morello G., Danielski C., Dickens D., Tremblin P., Lagage P. O., 2019, AJ, 157, 205
Morvan M., Nikolaou N., Tsiaras A., Waldmann I. P., 2020, AJ, 159, 109
Neil A. R., Rogers L. A., 2020, ApJ, 891, 12
Parmentier V., Crossfield I. J. M., 2018, in Deeg H., Belmonte J., eds, Handbook of Exoplanets. Springer, Cham, Switzerland, p. 1
Sarkis P., Mordasini C., Henning T., Marleau G. D., Mollière P., 2021, A&A, 645, A79
Schwartz J. C., Cowan N. B., 2015, MNRAS, 449, 4192
Schwartz J. C., Kashner Z., Jovmir D., Cowan N. B., 2017, ApJ, 850, 154
Schwarz G., 1978, Ann. Stat., 6, 461
Sing D. K. et al., 2016, Nature, 529, 59
Stevenson K. B. et al., 2012, ApJ, 754, 136
Stevenson K. B. et al., 2017, AJ, 153, 68
Tamburo P., Mandell A., Deming D., Garhart E., 2018, AJ, 155, 221
Teske J. et al., 2021, ApJS, 256, 33
The Theano Development Team et al., 2016, preprint (arXiv:1605.02688)
Thorngren D. P., Fortney J. J., Lopez E. D., Berger T. A., Huber D., 2021, ApJ, 909, L16
Tinetti G. et al., 2018, Exp. Astron., 46, 135
Waldmann I. P., 2012, ApJ, 747, 12
Watanabe S., 2010, preprint (arXiv:1004.2316)
Wong I. et al., 2014, ApJ, 794, 134
Zhang M. et al., 2018, AJ, 155, 83

## APPENDIX A: ANALYSIS USING INFLATED UNCERTAINTIES

We also considered the brightness temperatures and uncertainties reported by Garhart et al. (2020). Their eclipse depths, eclipse depth uncertainties, and brightness temperatures are correct, but the brightness temperature uncertainties were derived by taking the relative uncertainty in eclipse depth and using that to calculate the uncertainty in brightness temperature (Deming, private communication). In their reanalysis, Baxter et al. (2020) used the eclipse depths and uncertainties reported by Garhart et al. (2020) but fully propagated those uncertainties through the Planck function, which is non-linear, to derive uncertainties on the brightness temperatures. Comparing the uncertainties reported in both works, the uncertainties of Garhart et al. (2020) are roughly twice as big as those reported by Baxter et al. (2020).

To see how our conclusions would change had we used the artificially inflated uncertainties, we refit the multiplanet hierarchical model from Section 3. According the ΔWAIC scores (Table A1), the wavelength-independent hierarchical model makes much better predictions than the non-hierarchical model, and marginally better predictions than the wavelength-dependent model. Again, we do not detect the trend of increasing 4.5–3.6 μm brightness temperature ratio, nor do we rule it out.

The refit brightness temperatures are shown in Figs A1 and A2. When measurement uncertainties are higher, Bayesian shrinkage is more dramatic.

**Table A1.** WAIC scores for the three models, using the artificially inflated brightness temperature uncertainties from Garhart et al. (2020). We also tabulate the standard error of the difference in WAIC between each model. The two hierarchical models both make equally good predictions, and significantly outperform the non-hierarchical model.

| Model | ΔWAIC | $\sigma_{\Delta\mathrm{WAIC}}$ |
|---|---|---|
| Wavelength dependent | 0.0 | 0.0 |
| Wavelength independent | 0.25 | 1.8 |
| Separate | 11.09 | 3.61 |

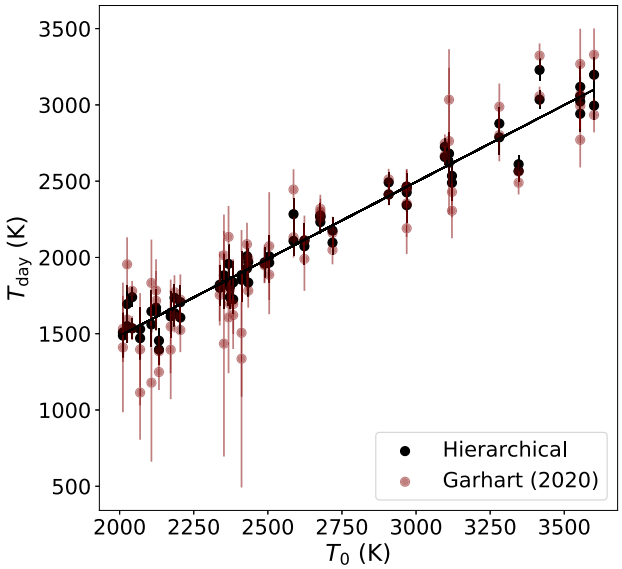


**Figure A1.** Dayside brightness temperatures refit with a wavelength-independent hierarchical model. We used the inflated uncertainties from Garhart et al. (2020). The red dots are the published values and the black line is the best-fitting trend line. The effects of Bayesian shrinkage are evident: the measurements are clustered closer to the line, and the uncertainties on the measurements are reduced.

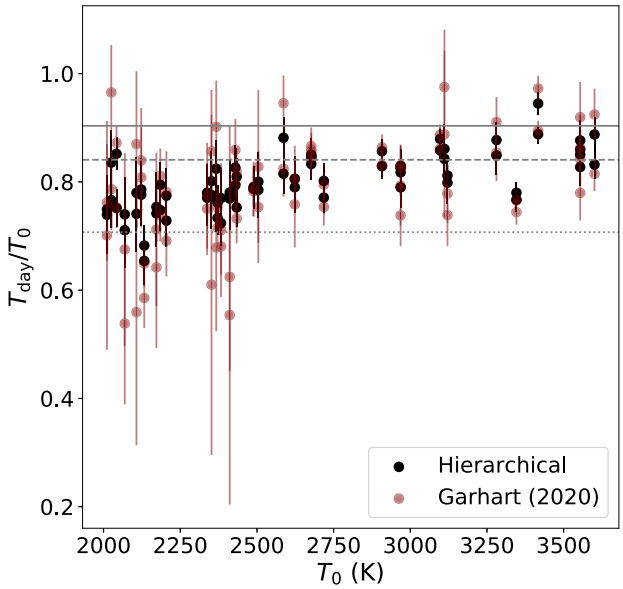


**Figure A2.** Dayside brightness temperatures from Fig. A1, this time scaled by irradiation temperature. The horizontal lines represent some theoretical limits on dayside temperature assuming a zero Bond albedo: the solid line assumes zero heat recirculation, the dashed line assumes a uniform dayside hemisphere but a temperature of zero on the nightside, and the dotted line assumes a uniform temperature at every location on the planet.

This paper has been typeset from a TEX/LATEX file prepared by the author.