

# TAU-REX I: A NEXT GENERATION RETRIEVAL CODE FOR EXOPLANETARY ATMOSPHERES

I. P. WALDMANN, G. TINETTI, M. ROCCHETTO, E. J. BARTON, S. N. YURCHENKO, AND J. TENNYSON  
Department of Physics & Astronomy, University College London, Gower Street, WC1E 6BT, UK; [ingo@star.ucl.ac.uk](mailto:ingo@star.ucl.ac.uk)

Received 2014 September 7; accepted 2015 January 26; published 2015 March 30

## ABSTRACT

Spectroscopy of exoplanetary atmospheres has become a well established method for the characterization of extrasolar planets. We here present a novel inverse retrieval code for exoplanetary atmospheres.  $\mathcal{T}$ -REx (Tau Retrieval for Exoplanets) is a line-by-line radiative transfer fully Bayesian retrieval framework.  $\mathcal{T}$ -REx includes the following features: (1) the optimized use of molecular line lists from the *ExoMol* project; (2) an unbiased atmospheric composition prior selection, through custom built pattern recognition software; (3) the use of two independent algorithms to fully sample the Bayesian likelihood space: nested sampling as well as a more classical Markov Chain Monte Carlo approach; (4) iterative Bayesian parameter and model selection using the full Bayesian Evidence as well as the Savage–Dickey Ratio for nested models; and (5) the ability to fully map very large parameter spaces through optimal code parallelization and scalability to cluster computing. In this publication we outline the  $\mathcal{T}$ -REx framework and demonstrate, using a theoretical hot-Jupiter transmission spectrum, the parameter retrieval and model selection. We investigate the impact of signal-to-noise ratio and spectral resolution on the retrievability of individual model parameters, both in terms of error bars on the temperature and molecular mixing ratios as well as its effect on the model’s global Bayesian evidence.

**Key words:** methods: data analysis – methods: statistical – radiative transfer – techniques: spectroscopic

## 1. INTRODUCTION

Remote sensing of atmospheres and inverse retrieval methods have a well established and long standing history. Beginning with pioneering work on our own Earth (e.g., Wark & Hilleary 1969; Conrath et al. 1970), we quickly extended our grasp to other planets in our solar system (e.g., Hanel et al. 1972, 1981; Conrath et al. 1973; Rodgers 1976). With the first detection of exoplanetary atmospheres (Charbonneau et al. 2002) we have taken this work beyond our solar system confines.

In recent years, the field of extrasolar spectroscopy has seen a increased effort in the development of data analysis and de-trending techniques (e.g., Swain et al. 2008, 2010; Carter & Winn 2009; Burke et al. 2010; Snellen et al. 2010; Thatte et al. 2010; Berta et al. 2012; Crouzet et al. 2012; Gibson et al. 2012; Waldmann et al. 2012, 2013; Waldmann 2012, 2014; Danielski et al. 2014; Kreidberg et al. 2014; Morello et al. 2014). With the maturation of these methodologies, we are obtaining a rapidly increasing number of exoplanetary emission and transmission spectra requiring interpretation. We refer the reader to Seager (2011) and Tinetti et al. (2013) and references within, for reviews of current spectroscopic results.

This ever increasing wealth of spectroscopic data of extrasolar planet atmospheres allows an unprecedented insight into the properties of these foreign worlds.

The interpretation of atmospheric spectra of extrasolar planets through *inverse atmospheric retrieval* modeling (e.g., Fletcher et al. 2007; Irwin et al. 2008; Terrien et al. 2008; Madhusudhan & Seager 2009; Lee et al. 2011; Benneke & Seager 2012; Line et al. 2012; Barstow et al. 2013; Griffith 2014) has become the industry standard. Line et al. (2013b) provide a recent and comprehensive review of currently existing exoplanetary atmospheric retrieval codes.

With greater accuracy in data often comes an increased complexity in its interpretation. In analogy to recent challenges in observational exoplanetary data analysis, one can identify

three major objectives for the interpretation of exoplanetary spectra.

**Sensitivity.** Given the often low resolution and low signal-to-noise ratio (S/N) of currently available exoplanetary spectra, an understanding of the limitations and degeneracies of spectroscopic models is paramount.

**Objectivity.** Are the results driven by model dependencies, over-constraint inputs or human biases? An idealized atmospheric retrieval should make no prior assumptions about the complex nature of exoplanetary atmospheres. While this is often infeasible, modern retrieval algorithms should be designed to take into account the broadest possible range of atmospheric models. It should then select among these models using a consistent and quantifiable metric of parameter and model adequacies.

**Big data.** With the increasing automation of exoplanet observations, the manual interpretation of atmospheric spectra will become infeasible. A modern retrieval algorithm should bear this in mind and allow for a high degree of intelligent automation and scalability to larger cluster computing.

In this paper, we introduce a new atmospheric retrieval code,  $\mathcal{T}$ -REx (Tau Retrieval for Exoplanets), which has been designed with the above objectives in mind. Here we will describe the overall architecture and atmospheric retrieval for transmission spectroscopy and dedicate a subsequent publication (I.P. Waldmann et al. 2015, in preparation) to the emission/reflection spectroscopy case and the parameterization of the temperature–pressure ( $T$ – $P$ ) profile.

### 1.1. $\mathcal{T}$ -REx

$\mathcal{T}$ -REx is a novel, fully Bayesian, retrieval code for exoplanetary atmospheres. In its current implementation,  $\mathcal{T}$ -REx includes the following features.

**Line-by-line.**  $\mathcal{T}$ -REx uses customized molecular and atomic line lists available directly from the *ExoMol*<sup>1</sup> project (Tennyson

<sup>1</sup> [www.exomol.com](http://www.exomol.com)

& Yurchenko 2012). In particular, *ExoMol* provides computed line lists valid over extended temperature ranges for a variety of molecules including water (Barber et al. 2006), ammonia (Yurchenko et al. 2011), methane (Yurchenko & Seager 2014), and a variety of diatomic molecules (Yadin et al. 2012; Barton et al. 2013, 2014; Barber et al. 2014). Besides line lists *ExoMol* provides cross sections (Hill et al. 2013) for *ExoMol* and other line lists. In this work cross sections for CO, NO, and CO<sub>2</sub> created from *HITEMP* (Rothman et al. 2010) and TiO from Schwenke (Schwenke 1998) are also used. Molecular (and atomic) absorption cross sections are calculated on an optimal linear or nonlinear wavelength grid resulting in a optimally sparse cross-section library with fine gridding of optically thick lines. This guarantees high computational efficiency without loss of accuracy.

Non-parametric prior constraint. The priors to the Bayesian retrieval, such as number and type of molecules considered, abundance, and temperature ranges, are not manually set by individual users but automatically determined by  $\mathcal{T}$ -REx based on the probability of individual molecules being present in the exoplanetary spectrum. The *Marple* module is a custom built pattern recognition package capable of rapidly identifying likely absorbers/emitters in the exoplanetary spectra from large line list archives (e.g., *ExoMol*, Tennyson & Yurchenko 2012; *HITRAN*<sup>2</sup>, Rothman et al. 2009, 2013; and *HITEMP*, Rothman et al. 2010). Such an approach is highly efficient as a very large number of molecules/atoms/ions can be considered and minimizes the human bias in selecting “key atmospheric components.”

Bayesian Model Selection. The code can be run to integrate over the full likelihood space of the Bayesian argument allowing for the Bayesian partition function, also called the Bayesian Evidence, to be calculated. This allows for the posterior distributions of model parameters to be calculated, as well as the adequacy of the model itself, given the data, to be assessed and iteratively optimized.

Scalability. Even the simplest retrieval cases can feature a high dimensional likelihood space. By relying on nested sampling (NS) approaches, we can naturally achieve an excellent multi-core processor scalability and full parallelization of the code, allowing us to fully map possible correlations in the likelihood space.

## 2. CODE OVERVIEW

The retrieval code discussed here is based on a fully modular, object oriented architecture. Among others, one of the main advantages of a modular approach is a more flexible and structured approach to complex programs. Throughout this paper we will follow this modular approach in describing  $\mathcal{T}$ -REx’s individual components.

The  $\mathcal{T}$ -REx design is illustrated in Figure 1 and can be broadly subdivided into four main programmatic segments:

1. *Inputs* (Sections 3 and 5)—these include global parameters, the observed exoplanetary spectrum and the absorption cross sections for the range of molecules to be considered. The *Marple* module (point 1 in Figure 1) acts as extra input to the main code, providing a best

initial guess of the atmospheric composition of the extrasolar planet to  $\mathcal{T}$ -REx.

2. *Model and data handling*—defines the radiative transfer forward model, the TP profile, overall input data handling. The Central Data Module (point 2 in Figure 1) acts as an abstraction layer between model, minimization, and data, allowing for an easy interchangeability of models, minimization/retrieval techniques and data types.
3. *Retrieval* (Sections 6)—contains three minimization codes: (1) Limited-memory Broyden–Fletcher–Goldfarb–Shannon (LM-BFGS) algorithm, (2) Adaptive Markov Chain Monte Carlo (MCMC) sampling, and (3) Bayesian NS. For NS runs, the Bayesian Partition function is calculated and model selection is performed. Point 3 in Figure 1 illustrates that all these algorithms have a common standardized interface with the Central Data Module. This guarantees exact and comparable results for different retrieval techniques.

The *Occam* module (Section 7) performs Bayesian model selection on the outputs of the MCMC and NS. In the case of under or over-complete models, the *Occam* module updates the planetary transmission model in an iterative manner (point 4 in Figure 1).

4. *Output* (Section 8)—the final exoplanetary spectrum is returned along with all parameter posterior distributions, cross-correlations and Bayesian Evidence.

## 3. ATOMIC AND MOLECULAR LINE LISTS

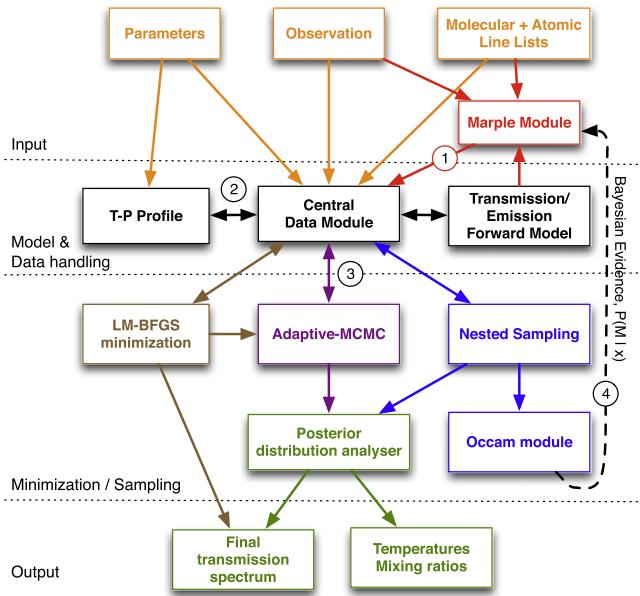
The optimal treatment of atomic and molecular line lists is key to the accuracy achieved by  $\mathcal{T}$ -REx. Throughout the code we perform line-by-line radiative transfer calculations at typically 50–100 times higher spectral resolution than the resolution of the observed spectrum to be analyzed. These “high-resolution” spectra are binned down (at each iteration of the code) to the data to calculate the  $\chi^2$  of the model fit. This ensures a correct treatment of optically thick absorption regions. In future versions of the code, we plan to include optimal nonlinear binning of line lists to further increase the computational efficiency without impacting model accuracies (E.J. Barton et al. 2015, in preparation).

$\mathcal{T}$ -REx allows for an easy and seamless inclusion of large numbers of line lists. For the scope of this paper we limit ourselves to line lists from absorption cross sections obtained from *ExoMol* but *HITRAN* line lists (or a combination of both) are equally natively supported. Automated pre-processing steps allow for conversions to a uniform data format with cross sections typically at  $\Delta\nu = 1.0 \text{ cm}^{-1}$  resolution. The cross-section library is generated at temperature intervals of 100 K (Hill et al. 2013) and upon execution of the main code interpolated to a user-set temperature resolution (typically 10 K). Two forms of cross section interpolation are available: (1) linear and (2) optimal. For the optimal case, we follow Hill et al. (2013) where the temperature interpolated cross-section  $\varsigma_{m,\lambda}(T)$  is given by

$$\varsigma_{m,\lambda}(T) = a_{m,\lambda} e^{-b_{m,\lambda}/T} \quad (1)$$

where  $m$  is the molecular/atomic species index,  $\lambda$  the wavelength,  $T$  the final temperature, and  $a$  and  $b$  are scaling

<sup>2</sup> [www.cfa.harvard.edu/hitran/](http://www.cfa.harvard.edu/hitran/)



**Figure 1.** Flowchart illustrating the modular design of  $\mathcal{T}$ -REx. As described in the text,  $\mathcal{T}$ -REx is subdivided into four main parts: Input, Model and data handling, Retrieval/minimization and Output analysis.

factors given by

$$b_{m,\lambda} = \left( \frac{1}{T_2} - \frac{1}{T_1} \right)^{-1} \ln \frac{\varsigma_{m,\lambda}(T_1)}{\varsigma_{m,\lambda}(T_2)} \quad (2)$$

$$a_{m,\lambda} = \varsigma_{m,\lambda}(T_1) e^{b_{m,\lambda}/T_1} \quad (3)$$

where  $T_1$  and  $T_2$  are upper and lower temperatures, respectively.

#### 4. FORWARD MODEL

The transmission forward model is based on the Tau code by Hollis et al. (2013) but was optimized for a significantly higher computational efficiency. We will only give a brief summary of the transmission model and refer the interested reader to the relevant literature (e.g., Brown 2001; Lio 2002; Tinetti et al. 2012; Hollis et al. 2013). As previously mentioned, in this paper we will only describe the transmission part of  $\mathcal{T}$ -REx and an isothermal  $T$ - $P$  profile. We dedicate a second publication (I.P. Waldmann et al. 2015, in preparation) to a complete treatment of the emission case and  $T$ - $P$  profile parametrization.

The monochromatic intensity,  $I_\lambda(z)$ , of radiation passing through a gas is given by the *Beer–Bouguer–Lambert Law* as function of atmospheric altitude,  $z$ ,

$$I_\lambda(z) = I_\lambda(0) e^{-\tau_\lambda(z)} \quad (4)$$

where  $\lambda$  is the wavelength of the radiation,  $I_\lambda(0)$  the incident radiation intensity at the top of the atmosphere, and  $\tau_\lambda(z)$  the optical depth of the medium. For a given absorber,  $m$ , we can state the optical depth to be the integral of the absorption cross-section,  $\varsigma_m(\lambda)$ , the column density,  $\chi_m(z)$ , and the number

density,  $\rho_N(z)$ , over the optical path length  $l(z)$

$$\tau_{\lambda,m}(z) = 2 \int_0^{l(z)} \varsigma_m(\lambda) \chi_m(z) \rho_N(z) dl \quad (5)$$

where the path length is dependent on the geometry of the transmission through the planet's terminator (see Figure 2 in Hollis et al. 2013). The overall optical depth is now given by the sum of the individual optical depths

$$\tau_\lambda(z) = \sum_{m=1}^{N_m} \tau_{\lambda,m}(z) \quad (6)$$

where  $N_m$  is the total number of absorbing species,  $m$ . We can now calculate the equivalent atmospheric depth,  $\alpha_\lambda$ , by summing over all atmospheric depth layers,  $z$ ,

$$\alpha_\lambda = 2 \int_0^{z_{\max}} (R_p + z) (1 - e^{-\tau_\lambda(z)}) dz \quad (7)$$

where  $z_{\max}$  is the maximum depth of the atmosphere considered. The total transit depth as a function of  $\lambda$  is hence given by

$$\delta_\lambda = \frac{R_p^2 + \alpha_\lambda}{R_*^2} \quad (8)$$

where  $R_p$  and  $R_*$  are the radii of the planet and star, respectively.

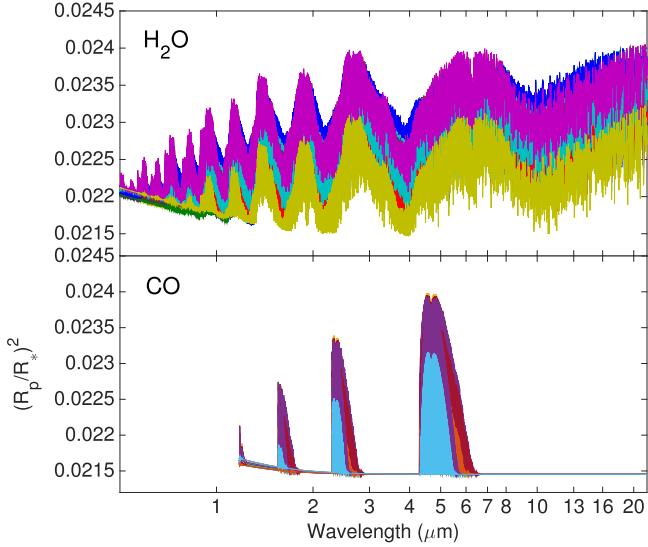
$\mathcal{T}$ -REx provides a full implementation of Rayleigh and Mie scattered as well as cloudy atmospheres. We refer the reader to Hollis et al. (2013) for details of implementation. Retrieval degeneracies due to cloud models will be discussed in a separate publication (M. Rocchetto et al. 2015, in preparation).

#### 5. MARPLE MODULE

The purpose of the Marple module is to constrain the prior space of the Bayesian retrieval in an unbiased way. Given the unknown, varied and complex nature of exoplanet systems it is difficult to pre-suppose atmospheric compositions from “experience.” The most objective approach to atmospheric retrieval of exoplanets would be to assume no prior knowledge at all and to consider all combinations of all atmospheric absorbers known. While desirable, this is computationally infeasible due to the large number of free parameters and often limited spectral resolution of the observed data. The Marple module attempts to limit the number of possible absorbers by identifying likely molecules in the observed data using a pattern recognition algorithm. It attempts to identify absorption/emission features that are “typical” for a molecular/atomic species and computes the possibility of such indicative features pertaining to a specific molecule compared to all other options. In this sense, the algorithm is conceptually similar to well established facial recognition algorithms using “eigenfaces” (e.g., Turk & Pentland 1991; Cendrillon & Lovell 2000; Gevaert & de With 2013).

The algorithm is described in the following steps.

1. The Marple module generates a library of atmospheric spectra (Equation (7)) for each atmospheric species using the available absorption cross sections,  $\varsigma_{m,\lambda}(T)$ . For each species,  $m$ , spectra are produced for a large range of atmospheric temperatures,  $T$ , and mixing ratios  $\chi$  (typically  $1 \times 10^{-8} \leq \chi \leq 1 \times 10^{-1}$  and



**Figure 2.** Transmission spectra of  $\text{H}_2\text{O}$  (top) and CO (bottom) for temperature and abundance ranges of 600–2000 K and  $1 \times 10^{-5}$ – $1 \times 10^{-2}$  respectively. Planet/star and orbital parameters are taken to be similar to hot-Jupiter HD 209458b.

$500 \text{ K} \leq T \leq 2000 \text{ K}$ ). Figure 2 shows the transmission spectra of water and CO over a range of temperatures and compositions. Where the molecular absorptions are strongest so are the variations. Note that a temperature range can be set for computational efficiency.

2. Characteristic spectral features for an atmospheric species vary significantly over the temperature and mixing ratio ranges computed above. These features are key to the identification of the absorber/emitter.

We can capture these significant variations using a principal component analysis (PCA, Jolliffe 2007) where the first component typically indicates the wavelength range over which these features are most prominent (i.e., the amplitude of the variation) and the second component reflects the modulation on the bulk variation, i.e., the features' morphology. Hence, for each atmospheric species, we compute the first and second principal components (PCs) over the range of spectra produced above using a single-value-decomposition (SVD).

The SVD of the column vector of spectra  $\alpha$  is given by

$$\alpha_m(T, \chi) = U \Sigma V^T |_m \quad (9)$$

where  $U$  and  $V$  are the left and right unitary matrices, respectively, and  $\Sigma$  is the diagonal eigenvalue matrix. Due to the large size of the matrices involved, we approximate Equation (9) with a randomized, truncated SVD algorithm (Halko et al. 2011; Martinsson et al. 2011). The individual PC is then given by

$$\begin{aligned} \mathbf{pc}_{n,m} &= U_n \Sigma_n |_m \\ &= \alpha_m(T, \chi) V_{n,m} \end{aligned} \quad (10)$$

where  $n$  is the PC index. Figure 3 shows the first (blue) and second (red) PCs of water. In this case spectral features are preserved in both components.

3. We now use the first PC calculated above to create the “feature mask,”  $\psi_m(\lambda)$ , for a given species. This masking guarantees that only wavelengths regions where a given

molecule absorbs/emits are correlated against the observed spectrum. The feature mask is given by

$$\psi_m(\lambda) = \begin{cases} 1 & \text{if } \frac{\mathbf{pc}_{1,m} - \arg \min(\mathbf{pc}_{1,m})}{\arg \max(\mathbf{pc}_{1,m})} > \eta \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

where “ $\arg \min$ ” and “ $\arg \max$ ” stand for the minimal and maximal values of the argument or array. This boolean mask identifies at which wavelengths the characteristic spectral features are stronger than the threshold parameter  $\eta$ . In other words, the  $\eta$  parameter sets the threshold between “molecule present” and “molecule absent” over a given wavelength range. We find  $\eta = 0.2$  to be a good choice for data with broad wavelength coverage. For spectra consisting of very few (<20) data points over a narrow wavelength range, the user may not want to exclude (i.e., mask) any data points from the analysis. In such cases setting  $\eta$  to a low value, e.g.,  $\eta = 0.05$ , will effectively prevent any wavelength range masking. A range of Marple module sensitivities can be explored by leaving  $\eta$  as free parameter, within user specified limits, over which the Occam module (Section 7) can iterate.

Figure 4 shows the creation of the “feature mask” for water and CO. Wavelengths with the normalized  $\mathbf{pc}_{1,m}$  bigger than  $\eta$  will be included into the feature mask of molecule  $m$ .

4. For each species we now convolve the second PC and the observed data  $x$  with the “feature mask” to obtain the masked PC and observed data vectors  $\widehat{\mathbf{pc}}_{2,m}$  and  $\widehat{x}_m$ , respectively,

$$\widehat{\mathbf{pc}}_{2,m} = \psi_m \otimes \mathbf{pc}_{2,m} \quad (12)$$

$$\widehat{x}_m = \psi_m \otimes \bar{x} \quad (13)$$

where  $\otimes$  denotes the convolution operator and  $\bar{x}$  is the normalized observed data vector given by

$$\bar{x} = \left( \frac{x - \arg \min(x)}{\arg \max(x)} \right). \quad (14)$$

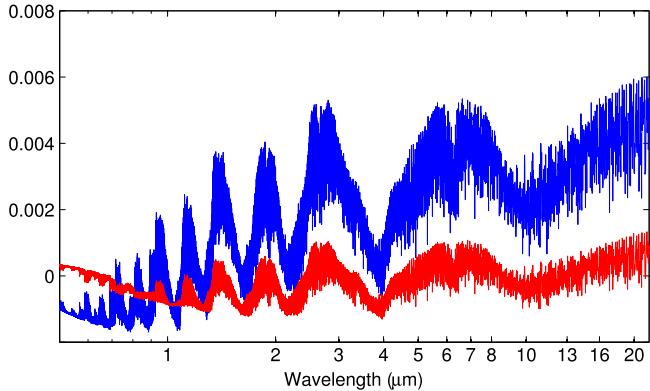
5. In order to select the set of best matching molecular/atomic species to the observed spectrum, we implement a variant of the K-nearest-neighbor (Cover & Hart 1967; Altman 1992) algorithm based on the euclidian distance between the spectral library PC vectors and the observed data. For this we calculate the  $L^2$  norm (also known as Euclidean norm or Euclidean distance),  $\mathfrak{d}_m$ , between masked data and the masked second PC for each molecule,  $m$

$$\mathfrak{d}_m = \frac{1}{N} \| \widehat{x}_m - \widehat{\mathbf{pc}}_{2,m} \|_2 \quad (15)$$

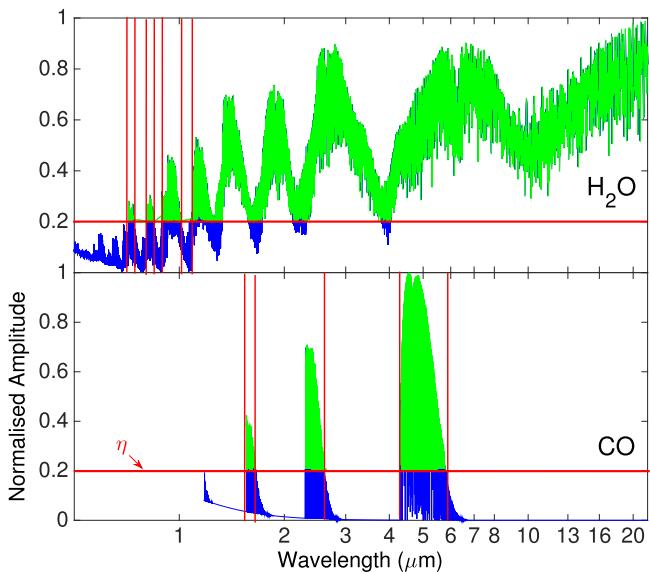
where  $N$  is the total number of data points in  $x$ . We now sort the euclidian distances in ascending order to form the monotonically increasing sequence of  $\mathfrak{d}_m$

$$f(\mathfrak{d}_m) = \{ \mathfrak{d}_{1,m}, \mathfrak{d}_{2,m}, \dots, \mathfrak{d}_{\phi,m} \} \quad (16)$$

where  $\phi$  is the sequence index and  $|\mathfrak{d}_{\phi-1}| < |\mathfrak{d}_\phi|$ . The total distance is given by  $\mathfrak{d}_{\text{total}} = \sum_\phi^M \{ \mathfrak{d}_\phi \}$ , where  $M$  is the total number of molecules considered. The algorithm



**Figure 3.** First (blue) and second (red) principal components of the water transmission spectrum library shown in Figure 2. The first component is used to create the “feature mask,” see text, whereas the second component is correlated against the observed data to identify possible matches between the observed spectra and water features.



**Figure 4.** Showing the creation of the “feature mask,”  $\psi_m(\lambda)$ , for  $\text{H}_2\text{O}$  (top) and  $\text{CO}$  (bottom). In blue are the first principal components of the molecules while in green are the spectral features selected for the mask. The horizontal red line indicates the threshold parameter  $\eta$ . Equation (11). Spectral features above this threshold are taken to be “significant features of the absorber” and included in the feature mask. The vertical red lines show the major cuts in the feature mask. For the case of  $\text{H}_2\text{O}$ , being an absorber/emitter across a broad wavelength range, most wavelengths are included in the feature mask  $\psi_{\text{H}_2\text{O}}(\lambda)$ .  $\text{CO}$  only absorbs/emits in discrete wavelength ranges and only those will be included in the feature mask of  $\text{CO}$ ,  $\psi_{\text{CO}}(\lambda)$ .

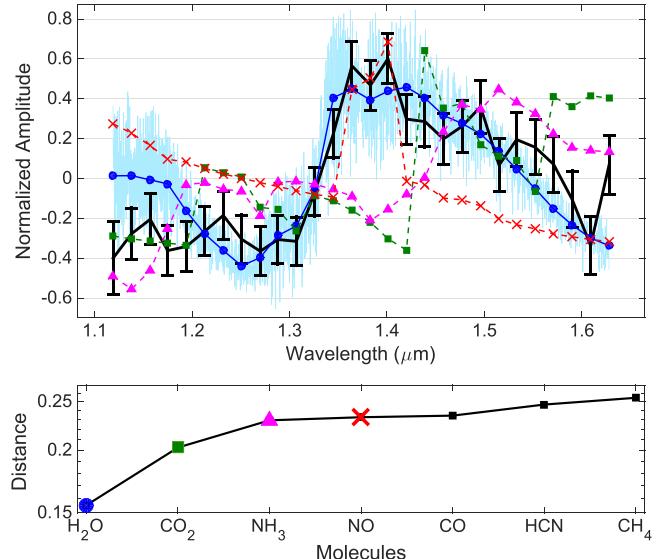
distinguishes between the cluster of best matching (low  $\delta$ ) and worst matching species (high  $\delta$ ) by finding the series index associated with the highest second derivative of  $f(\delta_m)$

$$\varphi = \phi \text{ where } \left[ \arg \max \left( \frac{d^2 f}{d \delta^2} \right) \right] \quad (17)$$

the number of molecules selected by the preprocessor,  $N_m$ , is then given by

$$N_m = \begin{cases} \varphi & \text{if } \varphi > N_{m,\min} \\ N_{m,\min} & \text{otherwise} \end{cases} \quad (18)$$

where  $N_{m,\min}$  is a minimal number of molecules to be



**Figure 5.** Top: Hubble/WFC3 transmission spectrum of HD 209458b (Deming et al. 2013; black). Overplotted are the principal components of the four best matching molecules binned to the resolution of the data:  $\text{H}_2\text{O}$  (blue dots),  $\text{CO}_2$  (green squares),  $\text{NH}_3$  (magenta triangles), and  $\text{NO}$  (red squares). The  $\text{H}_2\text{O}$  principal component is also shown at a resolution of  $R = 1000$  (light blue). Bottom: the normalized euclidean distance between each individual component and the data. It is clear that  $\text{H}_2\text{O}$  presents the best match to the data with other molecules being significantly worse.

selected which can be set by the user. The selected molecules are given by  $m_{\text{select}} = m_{\phi < \varphi}$ .

Once determined the Marple module passes its list of selected atmospheric species to the central data module (see Figure 1) which will prepare all inputs for further analysis. The efficiency of the Marple module is a function of spectral resolution and S/N of the data. This is self evident as any identification of features is impaired by a too coarse wavelength grid ( $R < 10$ ) or high noise levels ( $S/N < 5$ ). For those extreme cases, the user may specify a list of “must include” molecules in the parameter files to be considered by the  $\mathcal{T}$ -REx.

### 5.1. Example: HD 209458b

We demonstrate the Marple module using a transmission spectrum of the hot-Jupiter HD 209458b obtained by the *Hubble*/WFC3 camera (Deming et al. 2013). The top of Figure 5 shows the transmission spectrum in black and the PCs of the four best matching molecules. Note that all amplitudes are normalized and we only compare morphologies. The bottom panel summarizes the normalized Euclidean distances (Equation (15)) for individual molecules. Here the black continuous line represents the sequence  $f(\delta_m)$  in Equation (16). The Marple module returns water as the most likely molecule present with  $\text{CO}_2$  a more distant second. The presence of water as main absorbing species is in good agreement with the results of previous analyses (Deming et al. 2013; Madhusudhan et al. 2014).

## 6. RETRIEVAL

$\mathcal{T}$ -REx features three independent retrieval methods: (1) least-square minimization using a quasi-Newtonian LM-BFGS algorithm, Section 6.2, (2) an Adaptive, multi-chain MCMC

algorithm, Section 6.3.1, and (3) a NS algorithm using MultiNest, Section 6.3.2.

Programmatically, individual minimization routines submit standardized requests to the central data module which in turn handles all calls to the forward module, the  $T$ - $P$  profile, and required inputs (Figure 1). This modular approach guarantees consistency between model, data, and retrieval codes as well as a high degree of flexibility in the analysis of the observed data.

### 6.1. Prior Bounds

$\mathcal{T}$ -REx by default uses uniform priors for all free parameters. As default, the isothermal temperature bounds are  $T_{\text{equ}} \pm 200$  K, where  $T_{\text{equ}}$  is the planetary equilibrium temperature (this can either be derived by  $\mathcal{T}$ -REx given planetary/orbital parameters or set by the user). The molecular mixing ratios are bounded between  $0.0\text{--}1.0 \times 10^{-1}$  by default. All prior bounds can be manually specified by the user. The planet-star ratio,  $(R_p/R_*)^2$ , is treated as free parameter by default, with its upper/lower bounds derived from the reported observational uncertainty on this ratio. Griffith (2014) and Benneke & Seager (2013), among others, have noted strong degeneracies between the value of  $(R_p/R_*)^2$  and various retrieval parameters (e.g., H<sub>2</sub>O abundance and cloud opacities). We will further explore these degeneracies in a subsequent publication (M. Rocchetto et al. 2015, in preparation).

### 6.2. LM-BFGS Minimization

The least-square minimization allows us to obtain a quick look at the optimal model fit for the data without using the computationally expensive MCMC or NS routines. In this respect it is key to the pre-burning of the MCMC chain, as at least one chain can be started at the optimal solution and hence does not require a burn-in time (Brooks et al. 2011), as well as providing a valuable consistency check between model fits produced by the MCMC and MultiNest routines.

Large numbers of free parameters are often a limiting factor for simplex-downhill algorithms (e.g., Nelder & Mead 1965) commonly used. We find such amoeba algorithms to be insufficient and to often get stuck in local minima.  $\mathcal{T}$ -REx uses the LM-BFGS (Zhu et al. 1997; Morales & Nocedal 2011) algorithm instead, which being quasi-Newtonian uses the inverse Hessian matrix of the  $\chi^2$  surface to efficiently and robustly converge to the global maximum. We furthermore find the LM-BFGS to be more robust in the presence of observational noise than comparable methods.

### 6.3. Bayesian Analysis

The Bayesian argument is given by

$$P(\theta|x, \mathcal{M}) = \frac{P(x|\theta, \mathcal{M})P(\theta, \mathcal{M})}{P(x|\mathcal{M})} \quad (19)$$

where  $P(\theta, \mathcal{M})$  is the Bayesian prior which we take to be uniform throughout this paper. The number and type of absorbing species, as well as the equilibrium temperature of the planet defining the forward model,  $\mathcal{M}$ , are set by the Marple module (Section 5).  $P(\theta|x, \mathcal{M})$  is the posterior probability of the model parameters  $\theta$  given the data,  $x$  assuming the forward model  $\mathcal{M}$ . The likelihood,  $P(x|\theta, \mathcal{M})$  is

given by the Gaussian

$$P(x|\theta, \mathcal{M}) = \frac{1}{\varepsilon\sqrt{2\pi}} \exp\left[-\frac{1}{2}\sum_{\lambda}^N \left(\frac{x_{\lambda} - \mathcal{M}_{\lambda}}{\varepsilon_{\lambda}}\right)^2\right] \quad (20)$$

where  $\varepsilon$  is the error on the observed spectral point. As opposed to the NS described in the next section, an MCMC does not sample the Bayesian partition function (also known as Bayesian Evidence) and Equation (19) reduces to

$$P(\theta|x, \mathcal{M}) \propto P(x|\theta, \mathcal{M})P(\theta, \mathcal{M}). \quad (21)$$

#### 6.3.1. MCMC

MCMC routines are commonly used in the field of extrasolar planets (e.g., Braak 2006; Ford 2006; Bakos et al. 2007; Cameron et al. 2007; Knutson et al. 2007; Ter Braak & Vrugt 2008; Charbonneau et al. 2009; Bean et al. 2010; Burke et al. 2010; Goodman & Weare 2010; Gregory 2011; Kipping & Bakos 2011; Benneke & Seager 2012; Crouzet et al. 2012; Foreman-Mackey et al. 2013; Line et al. 2013b; Kreidberg et al. 2014; Madhusudhan et al. 2014).  $\mathcal{T}$ -REx provides an implementation of the Delayed-rejection (DR) Adaptive-MCMC (DRAM, Haario et al. 2006). We refer the interested reader to the cited literature and here only provide a brief overview. The DRAM algorithm differs from a more classical Metropolis–Hastings sampler (Metropolis & Rosenbluth 1953; Hastings 1970; Brooks et al. 2011) in two aspects: (1) It implements a DR algorithm and (2) an adaptive proposal distribution calibrated using the covariance of the sample path of the MCMC chain. For additional information on DRAM, we refer the reader to Appendix A and the relevant literature.

$\mathcal{T}$ -REx runs several MCMC chains in parallel to check convergence and increase the sampling of the likelihood space. The number of chains is user defined but usually set to 4–5 and limited by the number of available CPUs. The first primary chain is started at the optimal values determined by the LM-BFGS, avoiding significant burn-in time (Brooks et al. 2011). All secondary chains’ starting positions are offset from the optimum by a random distance and direction of at least 10% of the prior width. These secondary chains are run with a burn-in period of typically 10% of the total chain length. Burn-in and chain lengths are user defined.

#### 6.3.2. Nested sampling

NS algorithms (Skilling 2004, 2006; Jasa & Xiang 2005; Mukherjee et al. 2006; Chopin & Robert 2010; Keeton 2011) are becoming increasing popular in extrasolar planets (e.g., Kipping et al. 2012; Placek et al. 2013) as well as Benneke & Seager (2013) for atmospheric retrieval. Here we include an implementation of MultiNest (Feroz & Hobson 2008; Feroz et al. 2009, 2013). MCMC algorithms are commonly used for parameter estimation by solving Equation (21). Whereas MCMC explores the likelihood space by means of a Markovian chain, NS performs a general Monte Carlo analysis which is periodically constrained by ellipsoids encompassing spaces of highest likelihoods. Note that unlike MCMC, NS does not depend on a pre-determined proposal density and can hence better explore highly degenerate and non-Gaussian regimes. Using NS, we can compute the Bayesian evidence (or simply evidence), which is given by the integral required to

normalizing Equation (21)

$$E = \int P(\theta | \mathcal{M}) P(x | \theta, \mathcal{M}) d\theta \quad (22)$$

where  $E = P(x | \mathcal{M})$  is the evidence. The evidence allows us to test the adequacy of the model itself and to perform model selection as described in the following section. Posterior distributions for parameter estimations are returned as by-product of MultiNest and should be similar to posteriors obtained by the MCMC. Note that through the very different sampling techniques and fewer constraints on the proposal density for NS, we expect MCMC posteriors to be a “smoothed” version of the NS’s.  $\mathcal{T}$ -REx allows the choice between importance nested sampling (INS) and the more classical NS. Through the sampling process, INS retrains all accepted as well as rejected proposal points which allows for a more accurate integration of the evidence (Feroz et al. 2013). NS (in its MultiNest implementation) is highly efficient and easily parallelizable, allowing an easy scaling to cluster computing. We here use the NS approach as our main means of retrieval with the MCMC implementation providing a valuable cross check on the final results.

## 7. MODEL SELECTION

For an inverse retrieval problem, such as the one discussed here, the idea of model selection is highly relevant but rarely discussed due to the computational expense and complexities involved. Notable examples of Bayesian model selection in atmospheric retrieval are Benneke & Seager (2013), Line et al. (2013a), and Swain et al. (2014). Here we explicitly make the distinction between optimal estimation of parameters and the adequacy of the parameter and/or model itself.

We perform two tests after each  $\mathcal{T}$ -REx run.

1. Parameter adequacy. Is a parameter (e.g., a given molecular species) required to describe the underlying physics? If not, is the model considered *over-complete*? In the case of *over-completeness* the forward model may be too complex (not obeying Occam’s razor). This can lead to overfitting in the worst case or in the best case a reduction in retrieval efficiency.
2. Model adequacy. Are parameters missing in the model considered, i.e., is the model *under-complete*? In the case of model *under-completeness* the data is better accounted for by a more complex model. For example, a cloudy exoplanetary atmosphere cannot be modelled adequately by a cloud-free atmospheric model. Here the presence of clouds could force a cloud-free model to compensate for the extra absorption using molecular/atomic absorbers. This introduced bias often cannot be discerned from parameter estimating algorithms such as MCMC, maximum likelihood and similar methods.

Determining a model that is adequate to the data’s complexity is hence paramount.

$\mathcal{T}$ -REx tries to perform model selection in an intelligent way through the Occam module. The Occam module will perform model selection until a complete model is determined. It will iterate through models, appropriately increasing or decreasing the model complexity through interaction with the Marple module (see Figure 1, point 4).

**Table 1**  
Jeffrey’s Scale for Model Selection

| 2ln(SDR)  | Preference for Simplified Model $\mathcal{M}_{\theta-\theta_\gamma}$ |
|-----------|--|
| >10       | Very strong preference for excluding $\theta_\gamma$                 |
| 10 to 6   | Strong preference for excluding $\theta_\gamma$                      |
| 6 to 2    | Substantial preference for excluding $\theta_\gamma$                 |
| 2 to 0    | Insignificant preference for excluding $\theta_\gamma$               |
|           | Preference for Complex Model $\mathcal{M}_\theta$                    |
| 0 to -2   | Insignificant preference for including $\theta_\gamma$               |
| -2 to -6  | Substantial preference for including $\theta_\gamma$                 |
| -6 to -10 | Strong preference for including $\theta_\gamma$                      |
| <-10      | Very strong preference for including $\theta_\gamma$                 |

### 7.1. Over-complete Models

The over-complete model features an unnecessary complexity. Here the desired model is a sub-set of the more complex model initially run. We refer to these models as being “nested.” Complexity in parametric models (such as the forward models of atmospheric retrieval) is usually synonymous with number of free-parameters. Hence we can define the nested model  $\mathcal{M}_{\theta-\theta_\gamma}$  as sub-set of the more complex one  $\mathcal{M}_\theta$ ,

$$\mathcal{M}_{\theta-\theta_\gamma} = \mathcal{M}_\theta \Big|_{\theta_\gamma=0} \quad (23)$$

where  $\theta$  is a column-vector of all model parameters and  $\theta_\gamma$  is an individual parameter. The Bayes factor (see Section 7.2) allows us to perform this model selection by marginalizing out individual parameters (Benneke & Seager 2013; Swain et al. 2014). For “nested” models we can derive the simpler to compute Savage–Dickey density ratio (SDR; Dickey 1971; Marin & Robert 2010; Verdinelli & Wasserman 2012; Verde et al. 2013)

$$SDR = \frac{P(\theta_\gamma | x, \mathcal{M}_\theta)}{P(\theta_\gamma | \mathcal{M}_\theta)} \Big|_{\theta_\gamma=0} \quad (24)$$

where  $P(\theta_\gamma | x, \mathcal{M}_\theta)$  is the marginalized posterior of  $\theta_\gamma$  and  $P(\theta_\gamma | \mathcal{M}_\theta)$  is its respective prior distribution. A comprehensive derivation and discussion of Equation (24) can be found in Verde et al. (2013). This ratio of densities at  $\theta_\gamma = 0$  is indicative of whether a simpler model not containing  $\theta_\gamma$  is sufficient or whether a more complex model is preferred by the data. To assess the significance of the evidence toward a complex rather than a simpler model, we compare the outcome of Equation (24) to the Jeffreys scale (Jeffreys 1961). We adopt a slightly modified version of Kass & Raftery (1995) in Table 1.

The Occam module calculates the Savage–Dickey ratio for each model parameter and adjust the model complexity accordingly in case of a strong preference for a simpler forward model.

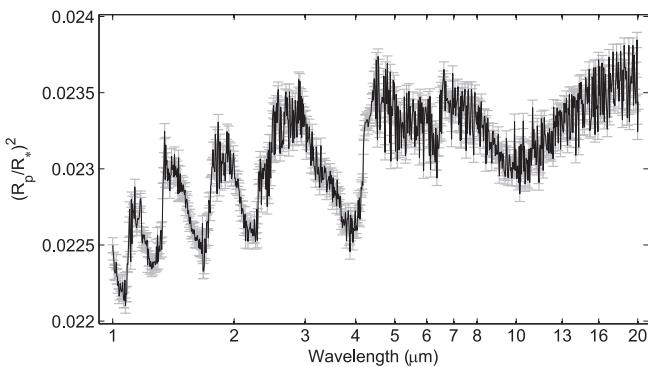
### 7.2. Under-complete Models

Should the model at hand not be over-complete, the Occam module tests for model under-completeness, i.e., whether the model is complex enough. For this we iteratively

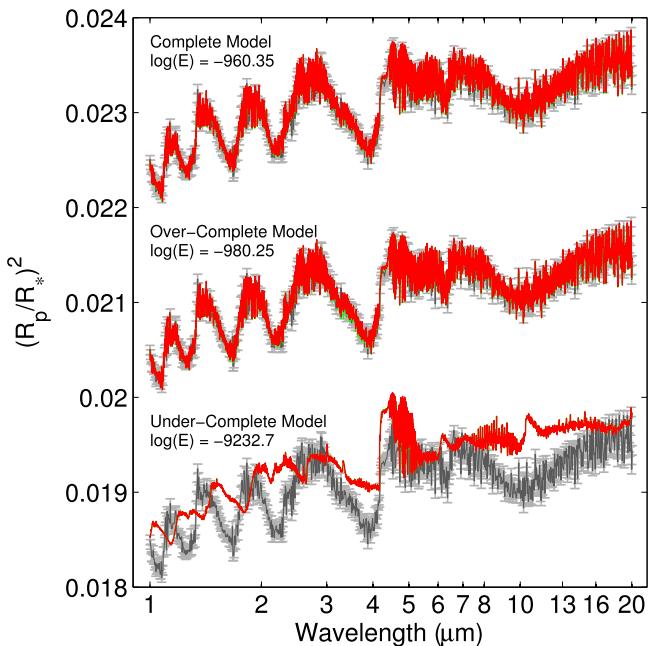
**Table 2**  
Model Input and Retrieval Results for LM-BFGS, MCMC and NS

| Parameters       | Model              | LM-BFGS               | MCMC  | Nested Sampling                               |
|------------------|--------------------|-----------------------|---|---|
| Temp. (K)        | 1400               | 1419.14               | $1403.30 \pm 9.88$                            | $1403.87 \pm 9.26$                            |
| H <sub>2</sub> O | $2 \times 10^{-3}$ | $1.94 \times 10^{-3}$ | $1.90 \times 10^{-3} \pm 3.27 \times 10^{-5}$ | $1.90 \times 10^{-3} \pm 3.11 \times 10^{-5}$ |
| CH <sub>4</sub>  | $2 \times 10^{-6}$ | $2.04 \times 10^{-6}$ | $2.25 \times 10^{-6} \pm 1.45 \times 10^{-6}$ | $2.17 \times 10^{-6} \pm 1.42 \times 10^{-6}$ |
| CO               | $2 \times 10^{-3}$ | $1.95 \times 10^{-3}$ | $1.97 \times 10^{-3} \pm 1.26 \times 10^{-4}$ | $1.97 \times 10^{-3} \pm 1.22 \times 10^{-4}$ |
| CO <sub>2</sub>  | $2 \times 10^{-5}$ | $2.41 \times 10^{-5}$ | $2.48 \times 10^{-5} \pm 2.59 \times 10^{-6}$ | $2.48 \times 10^{-5} \pm 2.60 \times 10^{-6}$ |
| NH <sub>3</sub>  | $2 \times 10^{-7}$ | $1.48 \times 10^{-6}$ | $1.18 \times 10^{-6} \pm 9.69 \times 10^{-7}$ | $1.18 \times 10^{-6} \pm 9.69 \times 10^{-7}$ |

**Note.** All values (but temperature) are in units of fractional column density.



**Figure 6.** Simulated example spectrum of a carbon-rich hot-Jupiter used in Section 9. Temperature and abundances of the main absorbers, H<sub>2</sub>O, CO, CO<sub>2</sub>, NH<sub>3</sub>, and CH<sub>4</sub> are given in Table 2. The bulk planet/star and orbital properties are based on the hot-Jupiter HD209458b.



**Figure 7.** Showing the best fitting models, in red, for the complete (top), over-complete (middle), and under-complete (bottom) model cases as described in Section 9. The best fitting models are offset along the ordinate for clarity and over-plotted on the “observed” spectrum in gray. The global Bayesian Evidences, log( $E$ ), are given for each case, quantifying the adequacy of each model given the data. As expected the evidence strongly favours the correct, complete model.

re-run the retrieval process allowing the Marple module to add the two next most likely molecular opacities to the current selection of opacities. We then compute the global model

evidence,  $E$ , and compute the Bayes factor (Kass & Raftery 1995; Weinberg 2012). The Bayes factor is given by the ratio of model probabilities  $P(\mathcal{M}_1|\theta)$

$$\frac{P(\mathcal{M}_2|\theta)}{P(\mathcal{M}_1|\theta)} = \frac{P(\mathcal{M}_2)}{P(\mathcal{M}_1)} \frac{P(x|\mathcal{M}_2)}{P(x|\mathcal{M}_1)} = \frac{P(\mathcal{M}_2)}{P(\mathcal{M}_1)} E_2 \quad (25)$$

which can be expressed as fraction of the Evidences and the prior distribution of the models. Most times we can assume the model priors to be identical  $P(\mathcal{M}_1) = P(\mathcal{M}_2)$ , reducing Equation (25) to

$$E_2/E_1 = \frac{P(x|\mathcal{M}_2)}{P(x|\mathcal{M}_1)}. \quad (26)$$

Using the Jeffrey’s scale the Occam module determines whether an improvement to the fit is achieved using a more complex model.

## 8. OUTPUTS

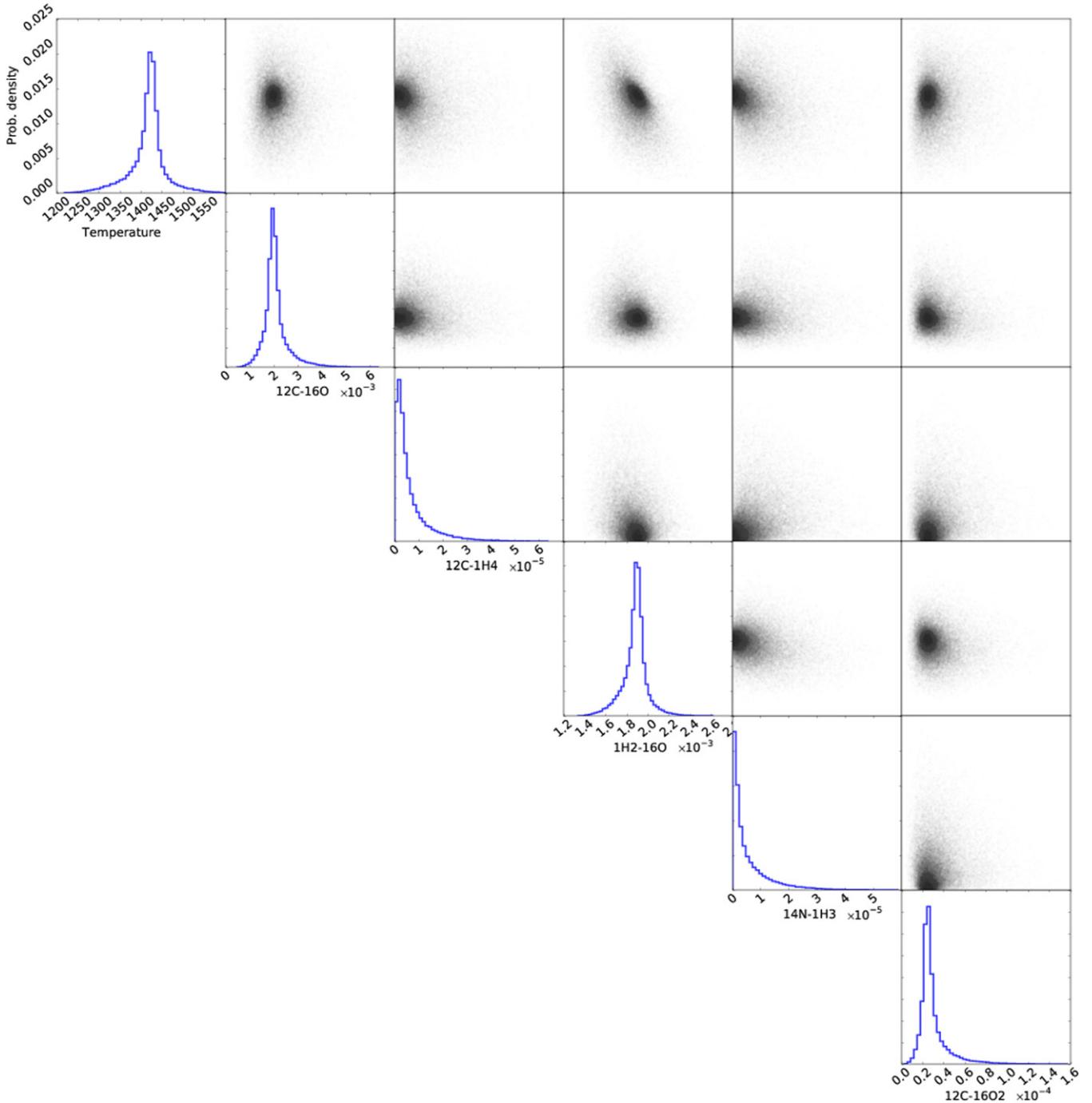
The output module generates the best-fit transmission model, plots of all marginalized and conditional posteriors, as well as statistics on individual parameters and model adequacy. Examples of these outputs can be found in the following section.

## 9. EXAMPLE

In this section we demonstrate the output of  $\mathcal{T}$ -REx using a simulated hot-Jupiter. We base the simulation on a HD209458b like planet/star system (Charbonneau et al. 2000; Southworth 2010) with temperature and bulk composition taken from Venot et al. (2014). We choose a wavelength range of 1–20  $\mu\text{m}$  at a constant resolution of  $R = 300$  and constant error bars of 50 ppm. Table 2 summarizes the inputs and Figure 6 shows the input spectrum to  $\mathcal{T}$ -REx. While such an example may be optimistic given currently available data we would like to note the following: (1) in order to demonstrate the retrieval accuracy of  $\mathcal{T}$ -REx one needs a precise data set, and (2) future observatories and missions (e.g., *James Webb Space Telescope*, E-ELT and dedicated missions) will yield data of comparable or better quality over broad wavelength ranges.

The data is passed through  $\mathcal{T}$ -REx as described in the previous sections. The Marple module suggested the correct molecules as potentially important absorbers given the data and their wavelength ranges. In addition to the molecules listed in Table 2, it also identified H<sub>2</sub>C<sub>2</sub> as possible absorber which was subsequently rejected by the Occam module and the transmission module was updated to reflect the true model of the data.

The retrieved temperature and abundance values for the LM-BFGS, MCMC, and NS algorithms are summarized in Table 2.

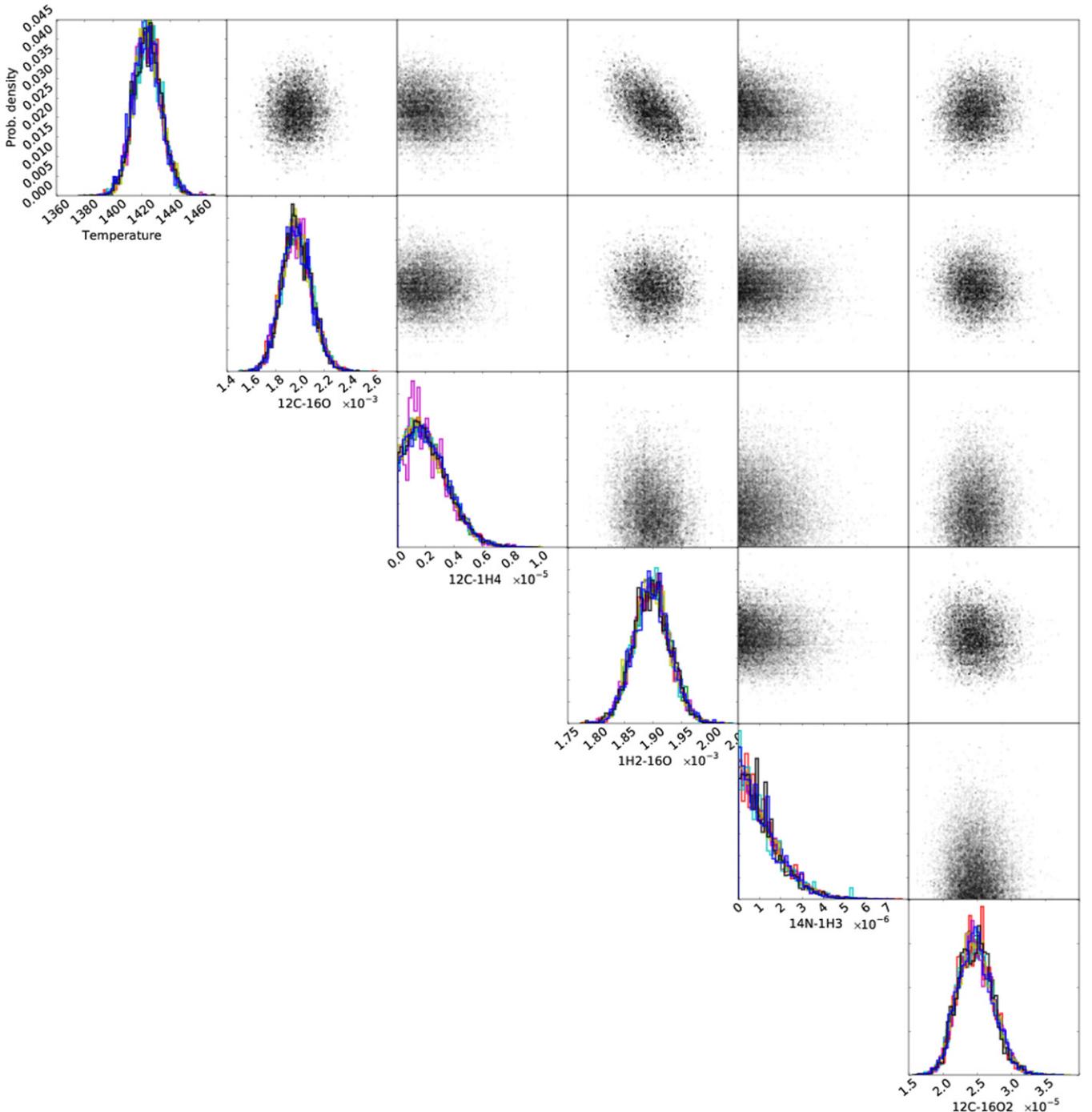


**Figure 8.** Showing the marginalized and conditional posterior distributions for the Nested Sampling for the complete model in Section 9. We find the highest correlation between atmospheric temperature and water absorption. With  $\text{H}_2\text{O}$  being the strongest absorber across the broadest wavelength range, this correlation between abundance and thermal broadening due to temperature changes is to be expected.

$\mathcal{T}$ -REx does not compute a formal error for the LM-BFGS result as only the MCMC and NS results are considered to be final data products. Figure 7 (top spectrum) shows the best-fit transmission model for the MCMC (green) and the NS (red) algorithms. Figures 8 and 9 show the marginalized and conditional posterior distributions for the MCMC and NS results, respectively. The MCMC results consist of 8 independent chains (note the different colors in the marginalized posteriors representing the results of individual chains) and  $2.5 \times 10^4$  samples each, including a 10% burn-in period. The NS results used 4000 initial live-points and  $8.3 \times 10^4$

replacements. Note the NS posteriors to be more kurtotic than the MCMC results. This is due to a finer sampling of the maximum likelihood space by the NS.

Table 2 summarizes the retrieved abundances. All major species in the simulated transmission spectrum as well as the isothermal temperature were retrieved with great fidelity by all retrieval methods. Mixing ratios for  $\text{NH}_3$  and  $\text{CH}_4$  were set purposefully low to test the retrievability of molecular abundances at the limits of data uncertainties. As shown in Section 9.1 and Table 3, these detections were identified as “insignificant” by the Occam module.



**Figure 9.** Showing the marginalized and conditional posterior distributions for the MCMC run of the complete model. Different colors in the marginalized posterior plots represent individual MCMC chains. The very good overlap of these independent chains indicates a good convergence of the code.

**Table 3**

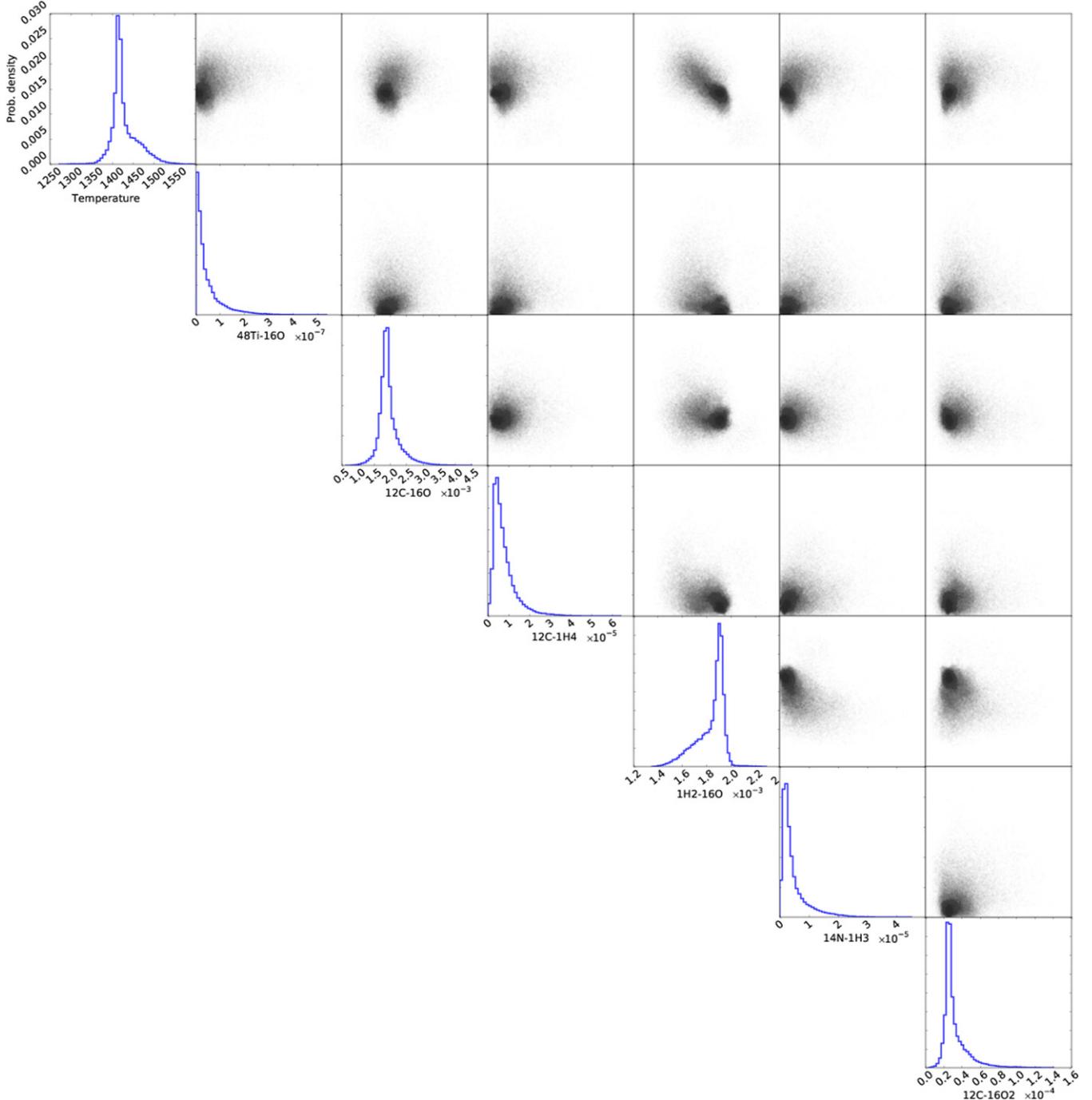
Savage–Dickey Density Ratio (SDR) for Overcomplete Model

| $2\ln(\text{SDR})$ | TiO  | CO    | $\text{CH}_4$ | $\text{H}_2\text{O}$ | $\text{NH}_3$ | $\text{CO}_2$ |
|--------------------|------|-------|---------------|----------------------|---------------|---------------|
| NS                 | 5.5  | -31.9 | -0.5          | -31.9                | 1.9           | -31.9         |
| MCMC               | 10.3 | -29.3 | 1.6           | -29.3                | 3.6           | -29.3         |

### 9.1. Model Selection

In addition to the complete model shown above, we have simulated an over-complete and under-complete model to test the model selection abilities of  $\mathcal{T}$ -REx. The over-complete

model contains TiO as additional absorber and the under-complete model lacks  $\text{H}_2\text{O}$ . Whereas in terms of  $\chi^2$  statistics we would expect a similarly valid fit for the over-complete model compared to the complete case (as the excessive parameters should converge to small values), we would expect a decrease in the overall model evidence as well as a clear discrimination of unnecessary complexity in the SDR. This behavior is indeed demonstrated by  $\mathcal{T}$ -REx. Figure 7 (middle) shows the model fit of the over-complete model and Figures 10 and 11 the posterior distributions of the NS and MCMC fits, respectively. As Figure 7 shows, the fit is maintained but at a

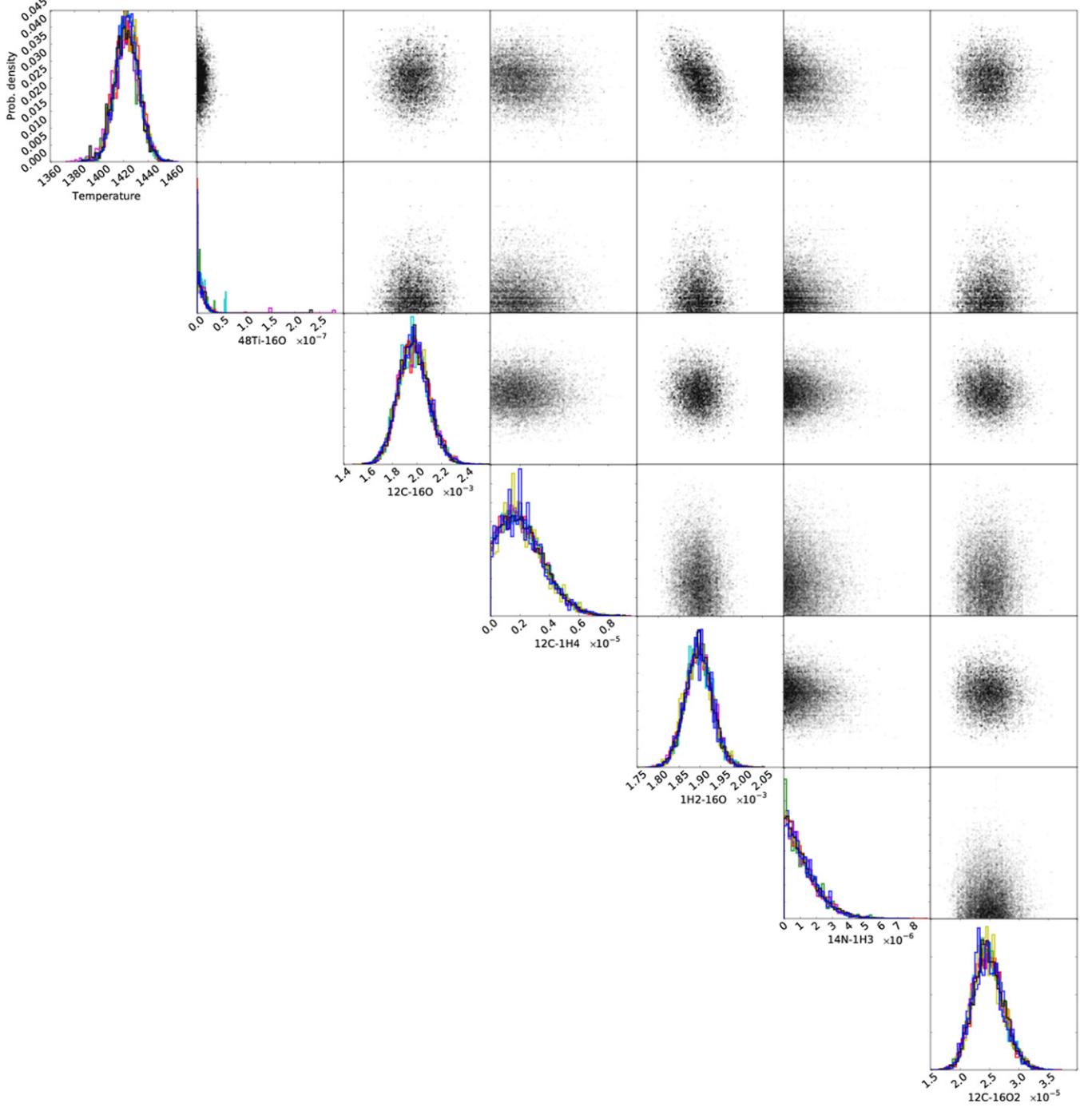


**Figure 10.** Showing marginalized and conditional posterior distributions of the Nested Sampling run for the over-complete model in Section 9.1. Otherwise identical to Figure 8.

lower global evidence,  $\log(E) = -980$  compared to  $\log(E) = -960$  for the correct model. On the Jeffrey's scale this results in a very strong preference for the overall simpler (i.e., complete) model. Table 3 shows the SDRs calculated from the NS and MCMC posteriors for  $\mathcal{M}_{\theta-\theta_\gamma}/\mathcal{M}_\theta$ , where  $\theta_\gamma$  is the molecule in question. The SDRs show a substantial to strong preference for the exclusion of TiO from the model and strongly confirm the inclusion of CO, H<sub>2</sub>O, and CO<sub>2</sub>. For the two low column density species, CH<sub>4</sub> and NH<sub>3</sub>, the SDRs neither include nor exclude either species but do not support a significant detection of the molecule in the data, as expected.

Differences in the SDR derived between NS and MCMC are due to the NS providing a tighter constraint on the marginalized posterior distributions than the MCMC, see Figures 10 and 11.

Figure 7 (bottom) shows the under-complete model excluding water. For under-complete models the  $\chi^2$  increases significantly as well as a very low global evidence of  $\log(E) = -9232$ . Figures 12 and 13 show the posterior distributions of the NS and MCMC runs, respectively. Both show a strong over dependence on high atmospheric temperatures, trying to fill in the missing opacities with an increased absorption due to an increased planetary scale height



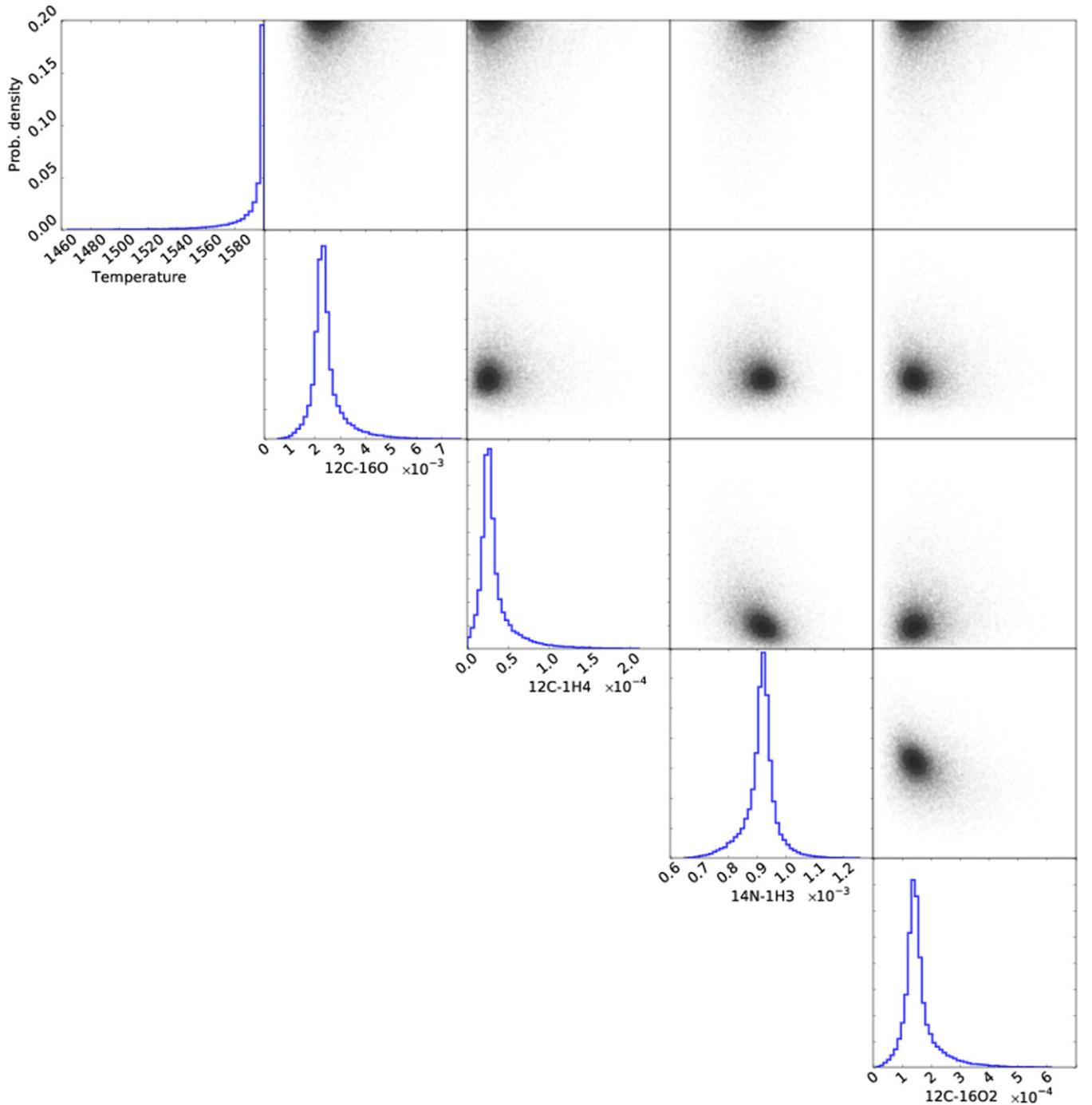
**Figure 11.** Showing marginalized and conditional posterior distributions of the MCMC run for the over-complete model in Section 9.1. Otherwise identical to Figure 9.

and an increased spectral broadening through the emergence of molecular hot-bands at higher temperatures.

### 9.2. Resolution and S/N

We now take the complete model from the previous section and reduce the S/N and resolution to explore the impact on the retrieval of exoplanetary spectra. Given the potentially large scope of such an exercise we here limit ourselves to a S/N-Resolution grid most representative of current data:  $R = 300, 200, 100, 50$ , and  $30$  and spectral error bars of  $\sigma = 10, 50, 100$ , and  $500$  ppm. Figure 14 shows the input data in red and the best

fitting transmission model at a resolution of  $R = 500$ . Whereas visually all spectra fit equally well, degeneracies between mixing ratios and temperature increase as resolution and S/N decrease. Whereas this result is intuitive, we find that the effect of degrading resolution and S/N is not uniform among parameters. Figure 15 shows the increase in error-bar (in percent) for the temperature posterior distribution derived. Here the reduction in S/N (i.e., increase in  $\sigma$ ) has a much more pronounced effect than the reduction in resolution, meaning that the planetary temperature can be derived with high confidence for low resolution data but not vice versa. Figure 16

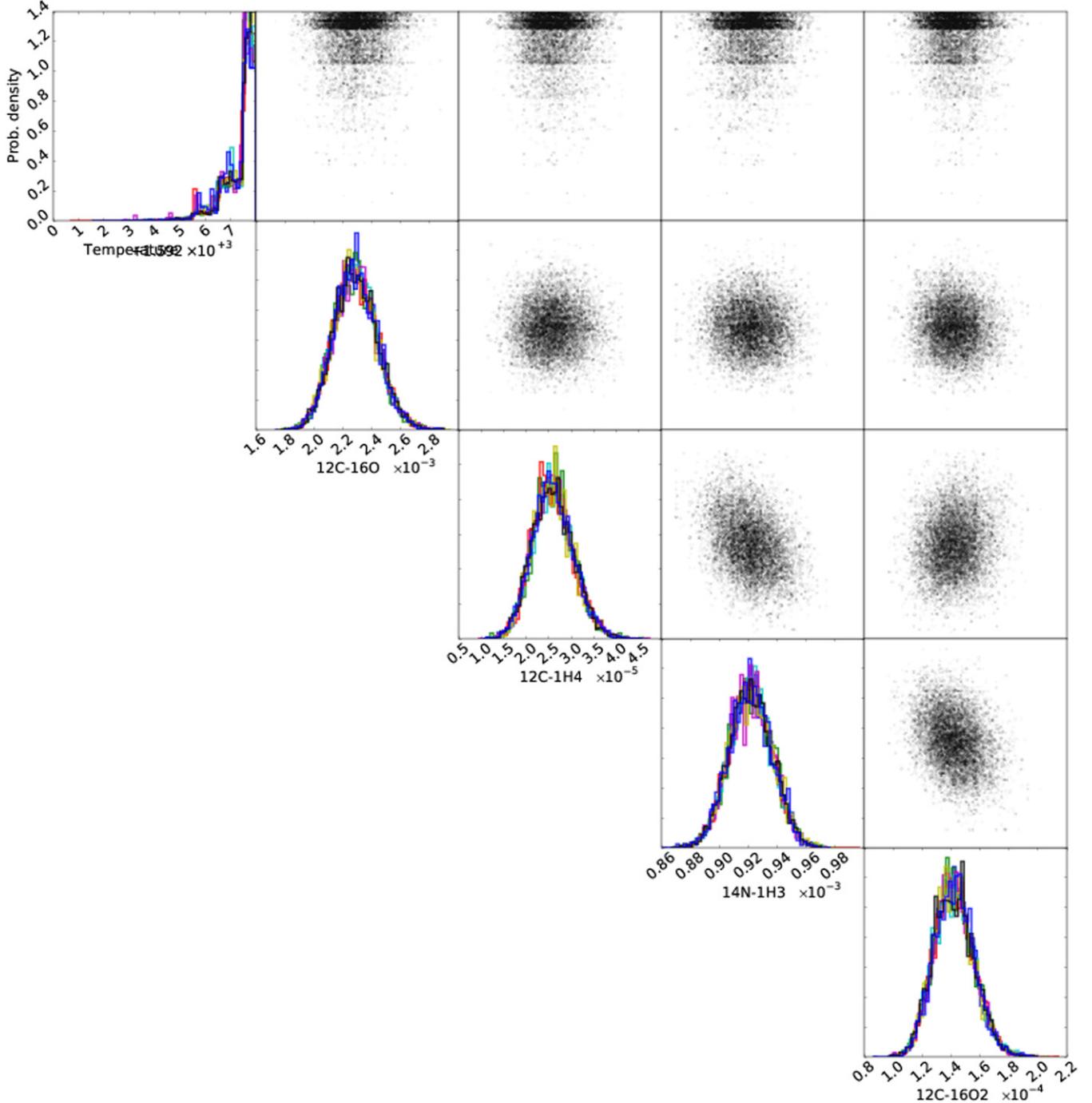


**Figure 12.** Showing marginalized and conditional posterior distributions of the Nested Sampling run for the under-complete model in Section 9.1. In the absence of the main absorbing species, H<sub>2</sub>O, T-REx tries to compensate for lacking opacity by increasing thermal broadening. This results in the planetary temperature converging to the upper end of the prior. Otherwise identical to Figure 8.

shows the same plot for the retrieval of the water mixing ratio. Water being very broad absorber in the NIR to mid-IR its abundance retrieval depends on resolution and S/N in approximately equal measures. Figure 17 plots the retrieval error bars of carbon-monoxide. With the spectral feature of CO being less broad than water we see that the dependence on S/N exceeds that of resolution (assuming that each CO feature is captured by at least one data point).

The loss of information through a reduced S/N is also demonstrated in Figure 18 showing the posterior distribution for all  $\sigma$  considered at  $R = 200$ . Here we can see the transition

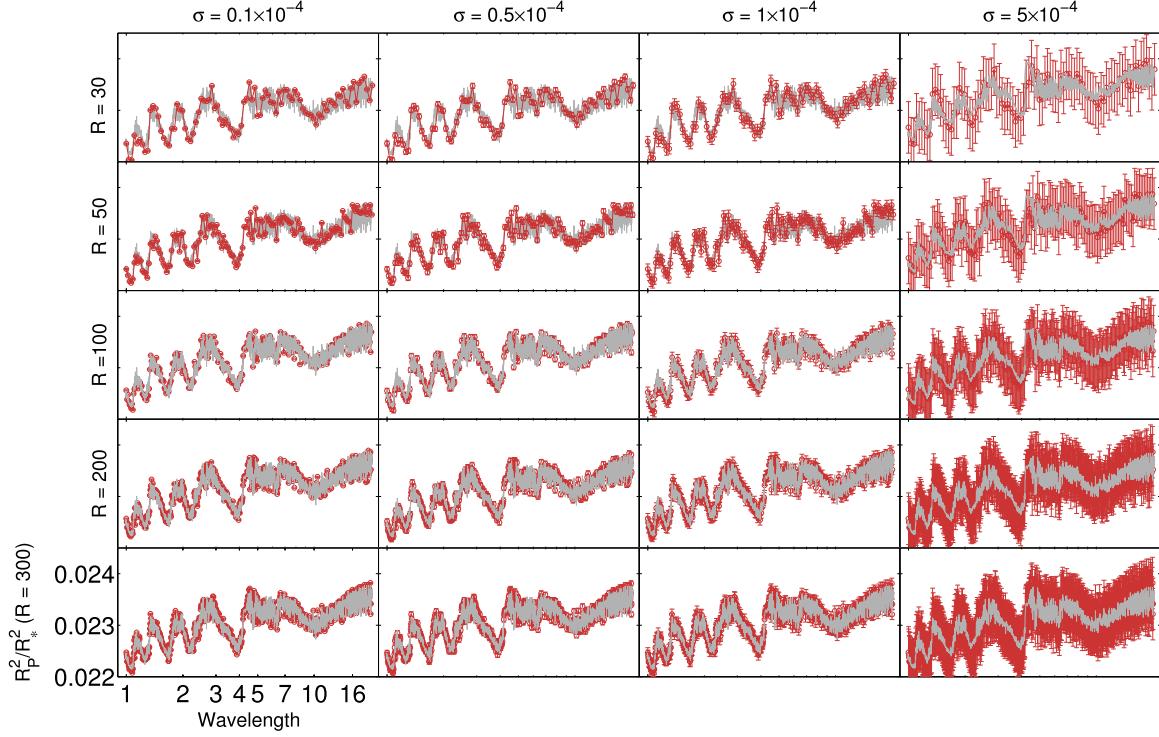
of the Bayesian argument going from data to prior dominated as the S/N decreases. Such a transition is gradual and we find a more regular occurrence of small local likelihood maxima as the likelihood surface “flattens out” at increasing  $\sigma$ . This is demonstrated by the the local maxima in the green curve ( $\sigma = 100$  ppm) compared to the best determined solution (blue curve,  $\sigma = 10$  ppm) and the prior driven solution (black curve,  $\sigma = 500$  ppm). The theoretical behavior of a Bayesian retrieval at low  $R$  and low S/N is an important result which will be address in detail in future work.



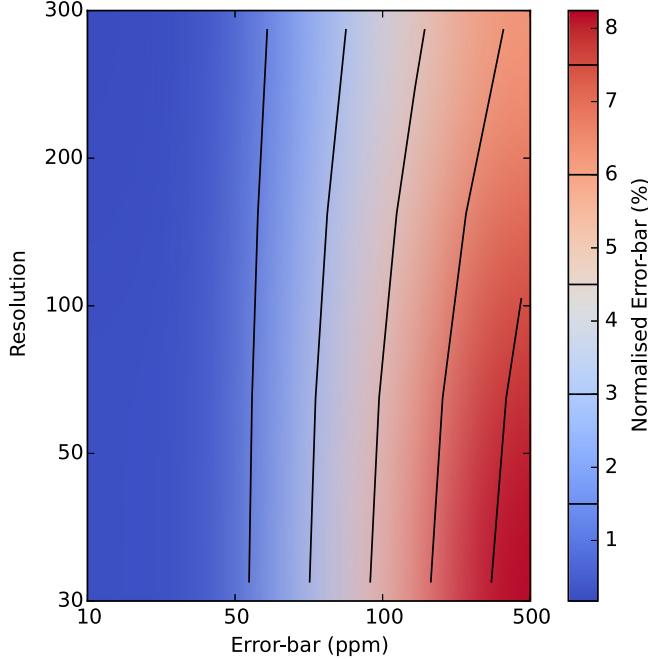
**Figure 13.** Showing marginalized and conditional posterior distributions of the MCMC run for the under-complete model in Section 9.1. Similar to Figure 12 the MCMC is converging to the upper temperature prior rapidly. This rapid convergence leaves “step” artefacts in the temperature posterior due to the discrete temperature resolution of 1 K. Step sizes can be set to an arbitrarily small value but in this case convergence to the upper prior bound is fast enough to only ever sample the top bins after the burn in period is completed. Otherwise identical to Figure 9.

Finally, we explore the effect of S/N and  $R$  on model selection. Similar to Table 3 we calculated the SDR for  $\text{H}_2\text{O}$ , CO and  $\text{NH}_3$  for the above S/N and  $R$  grid (Table 4). As with previous examples negative values below  $< -6$  indicate a strong detection with  $< -10$  being a decisive detection while equally positive values rule out a detection in the data. As expected, we see the detection evidence for  $\text{H}_2\text{O}$  and CO decrease at high  $\sigma$  and low  $R$ .  $\text{NH}_3$  (and  $\text{CH}_4$  not shown) remain undetected throughout as expected.

Line et al. (2012) present a complementary analysis for the emission spectroscopy case where the overall informational content, and the resulting possible number of retrievable free parameters, is calculated. Their analysis points at the inverse relationship between S/N and  $R$ . The lower the S/N of an observation, the higher the resolution must be to obtain the same degree of retrievability, and vice versa. This relationship we also find for the retrievability of individual parameters in the transmission case, e.g., CO in Figure 17.



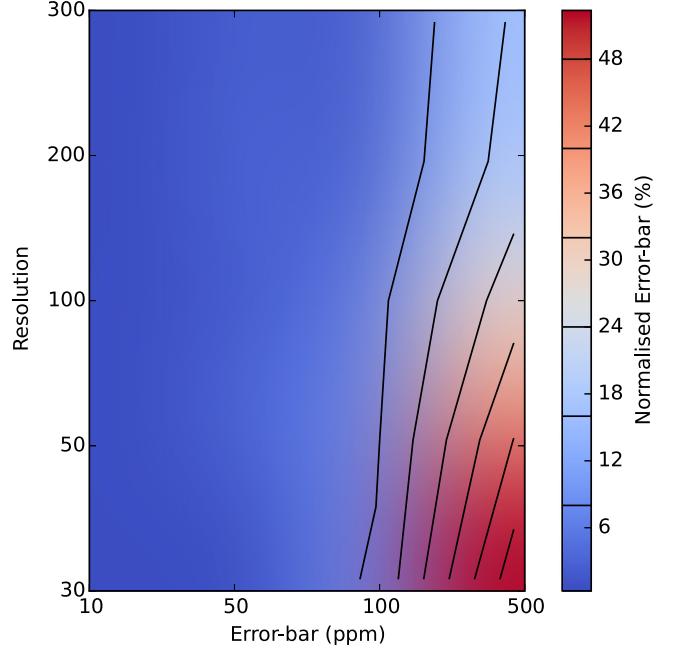
**Figure 14.** Best-fitting transmission model (gray) at  $R = 500$  superimposed on simulated input data (red) at resolutions  $R = 30, 50, 100, 200, 300$  and data-error bars  $\sigma = 10, 50, 100, 500$  ppm.



**Figure 15.** Temperature posterior standard deviation normalized by the ground-truth temperature (1400 K) as function of spectral resolution ( $R$ ) and the data-error bar. We find that the retrieval of the planetary temperature is dominated by the signal to noise of the data and less dominated by the resolution of the spectrum.

## 10. SUMMARY AND CONCLUSION

In this publication we have introduced the  $\mathcal{T}$ -REx retrieval code for exoplanetary atmospheres. As described in the introduction and shown throughout the text, we have based



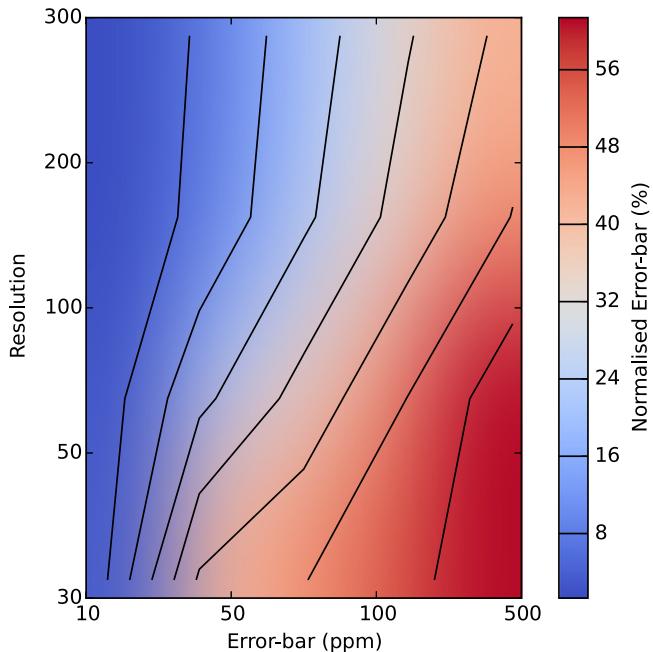
**Figure 16.**  $\text{H}_2\text{O}$  posterior standard deviation normalized by the ground-truth abundance ( $\chi_{(\text{H}_2\text{O})} = 2 \times 10^{-3}$ ) as function of spectral resolution ( $R$ ) and the data-error bar. The ability to retrieve water abundances remain relatively stable for high-resolution and low S/N data but significantly decreases as both S/N and  $R$  drop. Here posterior error-bars can reach the size of prior space.

the design of  $\mathcal{T}$ -REx on three guiding principles: (1) Sensitivity, (2) Objectivity, and (3) Big data.

$\mathcal{T}$ -REx incorporates a line-by-line radiative transfer code using state-of-the-art molecular opacity line lists by the *ExoMol* project. Atmospheric transmission models are run at

~50–100 times higher resolution than the observed data to ensure a correct treatment of optically thick absorption lines as well as allowing for a precise treatment of thermal line broadening through an arbitrarily finely sampled temperature grid.

Given the large number of potential absorbing/emitting species of an extrasolar planet, we have developed custom build pattern recognition software (the Marple module) to rapidly scan large molecular and atomic line-list archives for possible absorbing/emitting signatures in the observed spectrum. By not manually specifying a list of molecules “expected” to be present in the atmospheres of exoplanets we break potential human biases in the selection of the atmospheric model. In other words by not assuming anything about the atmospheres composition and structure we maximize the objectivity of the analysis from the start.



**Figure 17.** CO posterior standard deviation normalized by the ground-truth abundance ( $\chi_{\text{CO}} = 2 \times 10^{-3}$ ) as function of spectral resolution ( $R$ ) and the data-error bar. The CO retrieval more strongly depends on S/N other than for the resolution grid considered here.

Whereas the Marple module sets the potential prior space of the fully Bayesian retrieval, the Occam module performs iterative Bayesian model selection and iteratively verifies the adequacy of individual parameters as well as the overall evidence of the atmospheric model itself.

By using efficient MCMC and NS techniques throughout, we are able to parallelize the sampling of the likelihood space making  $\mathcal{T}$ -REx natively scalable to cluster computing. This allows  $\mathcal{T}$ -REx to explore very large parameter spaces and accurately map correlation manifolds.

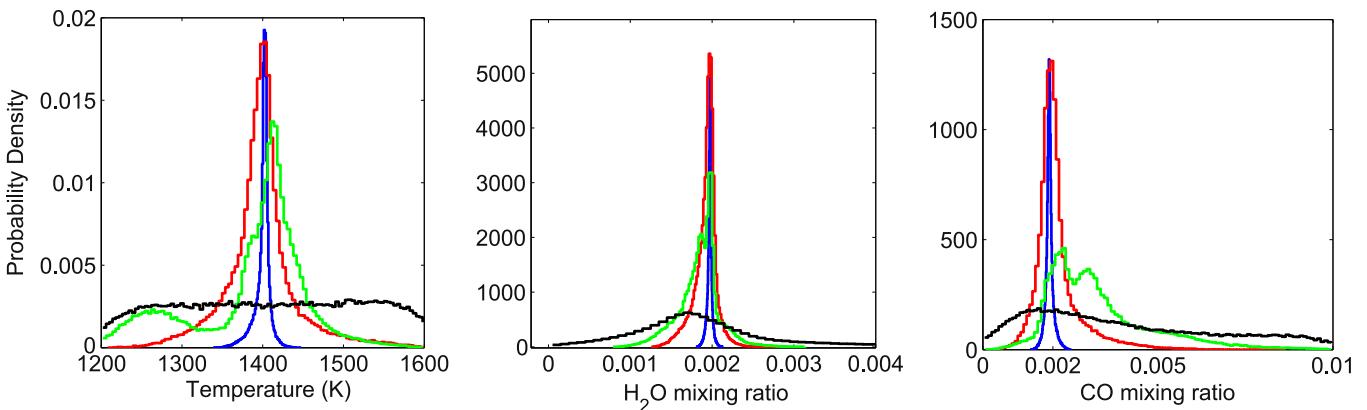
We demonstrated individual properties of  $\mathcal{T}$ -REx using a simulated hot-Jupiter and explored the model selection process of over-complete and under-complete models. The quality of the retrieval was investigated for varying resolutions and S/Ns of the input data and found to be consistent with expectations.

Future work will see a detailed treatment of emission spectroscopy in the framework of  $\mathcal{T}$ -REx, explore modeling degeneracies over large and short wavelength ranges and see the application to individual data sets.

**Table 4**  
Savage–Dickey Ratios (SDRs) for H<sub>2</sub>O, CO, and NH<sub>3</sub> for Data Error-bars of  $\sigma = 10, 50, 100, 500$  ppm and Resolutions of  $R = 300, 200, 100, 50, 30$

| Resolution           | 10 ppm | 50 ppm | 100 ppm | 500 ppm |
|----------------------|--------|--------|---------|---------|
| H <sub>2</sub> O 300 | −28.3  | −28.3  | −28.5   | −1.6    |
| 200                  | −28.3  | −28.6  | −28.5   | −2.6    |
| 100                  | −28.3  | −28.4  | −28.5   | −1.9    |
| 50                   | −28.2  | −28.2  | −28.6   | −0.8    |
| 30                   | −28.7  | −28.7  | −28.7   | −1.2    |
| CO 300               | −28.3  | −28.3  | −28.5   | −1.4    |
| 200                  | −28.3  | −28.6  | −28.5   | −1.0    |
| 100                  | −28.3  | −28.4  | −2.7    | −0.1    |
| 50                   | −28.2  | −28.2  | −2.3    | −0.1    |
| 30                   | −28.7  | −28.7  | −2.1    | 0.2     |
| NH <sub>3</sub> 300  | 8.0    | 5.7    | 4.8     | 9.1     |
| 200                  | 6.6    | 2.3    | 4.4     | 7.9     |
| 100                  | 0.5    | 3.0    | 3.6     | 6.6     |
| 50                   | −1.6   | 4.8    | 6.2     | 5.6     |
| 30                   | 7.7    | 7.7    | 8.4     | 5.5     |

**Note.** Negative values signify a detection of the molecule with values  $< -10$  being a very strong detection. Similarly, positive values  $> 6$  strongly indicate a non-detection. Values between  $-2$  and  $+2$  are inconclusive.



**Figure 18.** Three figures showing the posterior distributions for the planetary temperature (left), H<sub>2</sub>O (middle), and CO (right) mixing ratios at resolution  $R = 200$ . Colors represent data error bars. Blue:  $\sigma = 10$  ppm (scaled by a factor of 1/5 to improve comparability to other curves); red: 50 ppm; green: 100 ppm; black: 500 ppm.

With the maturation of data reduction techniques for exoplanetary spectroscopy we obtain higher and higher precision spectroscopy of these exotic atmospheres. With higher precision of the data often comes higher complexity in the interpretation. The goal of an ideal retrieval of atmospheric properties is to be able to capture said complexity while maintaining the highest possible degree of objectivity in the analysis.  $\mathcal{T}$ -REx presents a significant step toward this goal.

We thank the referee for providing useful comments. This work was supported by the ERC project numbers 617119 (ExoLights) and 267219 (ExoMol).

## APPENDIX A DRAM

In standard Metropolis–Hastings samplers each proposal step can either be accepted or rejected based on a fitness criterion and often a probability of acceptance when the fitness criterion is not met. Should the proposal be rejected the MCMC chain remains in the same position on the likelihood space. The DR mechanisms allows for a second (and third) proposal attempt to be made which is dependent on the previous chain as well as the previously rejected proposals. The adaptive proposal distribution based on its past history furthermore increases the efficiency and accuracy of the chain’s exploration as the proposal distribution is iteratively adapted to the target distribution. These features can be shown to significantly improve the efficiency of the MCMC chain in high dimensional likelihood space.

## APPENDIX B GLOSSARY

| Variable    | Description   | Equation Example |
|-------------|---|------------------|
| $N$         | Number of spectral points in data                               | ...              |
| $N_m$       | Number of molecules selected for retrieval                      | ...              |
| $m$         | Molecular species index   | ...              |
| $\lambda$   | Wavelength index  | ...              |
| $x$         | Data column vector  | ...              |
| $\bar{x}$   | Normalized data column vector                                   | 14               |
| $\varsigma$ | Absorption cross section  | 1                |
| $T$         | Temperature (K)   | ...              |
| $a, b$      | Absorption cross section temperature interpolation coefficients | 1, 2, 3          |
| $\chi$      | Atmospheric mixing ratio  | ...              |
| $\tau$      | Optical depth column vector over $\lambda$                      | ...              |
| $I(z)$      | Intensity column vector over $\lambda$ as function of $z$       | 4                |
| $z$         | Height in atmosphere  | ...              |
| $\alpha$    | Total atmospheric absorption column vector                      | 7                |

(Continued)

| Variable                   | Description   | Equation Example |
|----------------------------|---|------------------|
| $\alpha_m(T, \chi)$        | Total atmospheric absorption column vector as function of molecule, temperature, mixing ratio | 9                |
| $R_p$                      | Planetary radius  | ...              |
| $R_*$                      | Stellar radius  | ...              |
| $\delta$                   | Transit-depth column vector   | 8                |
| $U$                        | Left unitary matrix of single value decomposition   | 9, 10            |
| $\Sigma$                   | Diagonal matrix of single value decomposition   | 9, 10            |
| $V$                        | Right unitary matrix of single value decomposition  | 9, 10            |
| $pc$                       | Principal component vector  | 10               |
| $n$                        | Principal component index   |                  |
| $\psi_m$                   | Boolean data masking vector as function of molecule and wavelength                            | 11               |
| $\eta$                     | Molecule detection threshold coefficient  | 11               |
| $\hat{x}$                  | Masked data vector  | 13               |
| $\hat{pc}$                 | Masked PCA vector   | 12               |
| $\mathfrak{d}_m$           | $l_2$ -norm between normalized data and 2nd PC  | 15               |
| $f(\mathfrak{d}_m)$        | Monotonically increasing function of $\mathfrak{d}_m$   | 16               |
| $\phi$                     | Marple cluster index  | 16               |
| $\phi$                     | Marple cluster index for highest second derivative of $f(\mathfrak{d})$                       | 17               |
| $m_{\text{select}}$        | Marple determined molecular/atmoic species  | ...              |
| $\mathcal{M}$              | Exoplanet model   | ...              |
| $\theta_\gamma$            | Generic parameter of model $\mathcal{M}$  | ...              |
| $\theta$                   | Column vector of parameters of model $\mathcal{M}$  | ...              |
| $\sigma_\lambda$           | One sigma error at wavelength $\lambda$   | ...              |
| $P(\theta x, \mathcal{M})$ | Posterior probability distribution of $\theta$ given $x$ and $\mathcal{M}$                    | 19               |
| $P(x \theta, \mathcal{M})$ | Likelihood distribution of $\theta$   | 19               |
| $P(\theta, \mathcal{M})$   | Prior distribution of $\theta$  | 19               |
| $P(x \mathcal{M})$         | Bayesian partition function   | 19               |
| $E$                        | Bayesian Evidence   | 22               |

## REFERENCES

- Altman, N. S. 1992, *The American Statistician*, Vol. 46, 175
- Bakos, G. Á., Noyes, R. W., Kovacs, G., et al. 2007, *ApJ*, **656**, 552
- Barber, R. J., Strange, J. K., Hill, C., et al. 2014, *MNRAS*, **437**, 1828
- Barber, R. J., Tennyson, J., Harris, G. J., & Tolchenov, R. N. 2006, *yCat*, **6119**, 0
- Barstow, J. K., Aigrain, S., Irwin, P. G. J., Fletcher, L. N., & Lee, J. M. 2013, *MNRAS*, **434**, 2616
- Barton, E. J., Chiu, C., Golpayegani, S., et al. 2014, arXiv:1403.7952
- Barton, E. J., Yurchenko, S. N., & Seager, S. 2013, *MNRAS*, **434**, 1469
- Bean, J. L., Kempton, E. M.-R., & Homeier, D. 2010, *Natur*, **468**, 669
- Benneke, B., & Seager, S. 2012, *ApJ*, **753**, 100
- Benneke, B., & Seager, S. 2013, *ApJ*, **778**, 153
- Berta, Z. K., Charbonneau, D., Desert, J.-M., et al. 2012, *ApJ*, **747**, 35
- Braak, C. J. F. T. 2006, *Statistics and Computing*, **16**, 239
- Brooks, S., Gelman, A., Jones, G., & Meng, X. L. 2011, *Handbook of Markov Chain Monte Carlo* (London: Chapman and Hall)
- Brown, T. M. 2001, *ApJ*, **553**, 1006
- Burke, C. J., McCullough, P. R., Bergeron, L. E., et al. 2010, *ApJ*, **719**, 1796
- Cameron, A. C., Wilson, D. M., West, R. G., et al. 2007, *ApJ*, **380**, 1230
- Carter, J. A., & Winn, J. N. 2009, *ApJ*, **704**, 51
- Cendrillon, R., & Lovell, B. 2000, *Proc. SPIE*, **4067**, 269
- Charbonneau, D., Brown, T. M., Latham, D. W., & Mayor, M. 2000, *ApJL*, **529**, L45
- Charbonneau, D., Brown, T. M., Noyes, R. W., & Gilliland, R. L. 2002, *ApJ*, **568**, 377
- Charbonneau, D., Berta, Z. K., Irwin, J., et al. 2009, *Natur*, **462**, 891
- Chopin, N., & Robert, C. P. 2010, *Biometrika*, **97**, 741
- Conrath, B., Curran, R., Hanel, R., et al. 1973, *JGR*, **78**, 4267
- Conrath, B. J., Hanel, R. A., & Kunde, V. G. 1970, *JGR*, **75**, 5831
- Cover, T., & Hart, P. 1967, *ITIT*, **13**, 21
- Crouzet, N., McCullough, P. R., Burke, C., & Long, D. 2012, *ApJ*, **761**, 7
- Danielski, C., Deroo, P., Waldmann, I. P., et al. 2014, *ApJ*, **785**, 35
- Deming, D., Wilkins, A., McCullough, P., et al. 2013, *ApJ*, **774**, 95
- Dickey, J. M. 1971, *The Annals of Mathematical Statistics*, **42**, 204
- Feroz, F., Gair, J. R., Hobson, M. P., & Porter, E. K. 2009, *CQGra*, **26**, 215003
- Feroz, F., & Hobson, M. P. 2008, *MNRAS*, **384**, 449
- Feroz, F., Hobson, M. P., Cameron, E., & Pettitt, A. N. 2013, arXiv:1306.2144v2
- Fletcher, L. N., Irwin, P. G. J., Teanby, N. A., et al. 2007, *Icar*, **189**, 457
- Ford, E. B. 2006, *ApJ*, **642**, 505
- Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, *PASP*, **125**, 306
- Gevaert, W. J. R., & de With, P. H. N. 2013, *Proc. SPIE*, **8655**, 865503
- Gibson, N. P., Aigrain, S., Roberts, S., et al. 2012, *MNRAS*, **419**, 2683
- Goodman, J., & Weare, J. 2010, *Communications in Applied Mathematics and Computational Science*, **5**, 65
- Gregory, P. C. 2011, *MNRAS*, **410**, 94
- Griffith, C. A. 2014, *RSPTA*, **372**, 30086
- Haario, H., Laine, M., Mira, A., & Saksman, E. 2006, *Statistics and Computing*, **16**, 339
- Halko, N., Martinsson, P.-G., Shkolnisky, Y., & Tygert, M. 2011, *SIAM J. Sci. Comput.*, **33**, 2580
- Hanel, R., Conrath, B., Flasar, F. M., et al. 1981, *Sci*, **212**, 192
- Hanel, R. A., Conrath, B. J., Hovis, W. A., et al. 1972, *Sci*, **175**, 305
- Hastings, W. K. 1970, *Biometrika*, **57**, 97
- Hill, C., Yurchenko, S. N., & Seager, S. 2013, *Icar*, **226**, 1673
- Hollis, M. D. J., Tessenyi, M., & Tinetti, G. 2013, *CoPhC*, **184**, 2351
- Irwin, P. G. J., Teanby, N. A., de Kok, R., et al. 2008, *JQSRT*, **109**, 1136
- Jasa, T., & Xiang, N. 2005, in *AIP Conf. Proc.*, Vol. 803, Bayesian Inference and Maximum Entropy Methods in Science and Engineering, ed. K. H. Knuth et al. (New York: AIP), 189
- Jeffreys, H. 1961, *The Theory of Probability* (Oxford: Oxford Univ. Press)
- Jolliffe, I. T. 2007, *Principal Component Analysis* (2nd ed.; New York: Springer)
- Kass, R. E., & Raftery, A. E. 1995, *Journal of the American Statistical Association*, **90**, 773
- Keeton, C. R. 2011, *MNRAS*, **414**, 1418
- Kipping, D., & Bakos, G. 2011, *ApJ*, **733**, 36
- Kipping, D. M., Bakos, G. Á., Buchhave, L., Nesvorný, D., & Schmitt, A. 2012, *ApJ*, **750**, 115
- Knutson, H. A., Charbonneau, D., Allen, L. E., et al. 2007, *ApJ*, **447**, 183
- Kreidberg, L., Bean, J. L., Desert, J.-M., et al. 2014, *Natur*, **505**, 69
- Lee, J. M., Fletcher, L. N., & Irwin, P. G. J. 2011, *MNRAS*, **420**, 170
- Line, M. R., Knutson, H., Deming, D., Wilkins, A., & Desert, J.-M. 2013a, *ApJ*, **778**, 183
- Line, M. R., Zhang, X., Vasish, G., et al. 2012, *ApJ*, **749**, 93
- Line, M. R., Wolf, A. S., Zhang, X., et al. 2013b, *ApJ*, **775**, 137
- Liou, K. N. 2002, *An Introduction to Atmospheric Radiation* (New York: Academic Press)
- Madhusudhan, N., Crouzet, N., McCullough, P. R., Deming, D., & Hedges, C. 2014, *ApJL*, **791**, L9
- Madhusudhan, N., & Seager, S. 2009, *ApJ*, **707**, 24
- Marin, J. M., & Robert, C. P. 2010, *EJSta*, **4**, 643
- Martinsson, P.-G., Rokhlin, V., & Tygert, M. 2011, *Applied and Computational Harmonic Analysis*, **30**, 47
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. 1953, *JChP*, **21**, 1087
- Morales, J. L., & Nocedal, J. 2011, *ACM Transactions on Mathematical Software (TOMS)*, **38**, 7
- Morello, G., Waldmann, I. P., Tinetti, G., et al. 2014, *ApJ*, **786**, 22
- Mukherjee, P., Parkinson, D., & Liddle, A. R. 2006, *ApJ*, **638**, L51
- Nelder, J. A., & Mead, R. 1965, *CompJ*, **7**, 308
- Placek, B., Knuth, K. H., & Angerhausen, D. 2013, arXiv:1310.6764
- Rodgers, C. D. 1976, *RvGSP*, **14**, 609
- Rothman, L. S., Gordon, I. E., Barbe, A., et al. 2009, *JQSRT*, **110**, 533
- Rothman, L. S., Gordon, I. E., Barber, R. J., et al. 2010, *JQSRT*, **111**, 2139
- Rothman, L. S., Gordon, I. E., Babikov, Y., et al. 2013, *JQSRT*, **130**, 4
- Schwenke, D. W. 1998, *Chemistry and Physics of Molecules and Grains in Space*. Faraday Discussions No. 109, The Faraday Division of the Royal Society of Chemistry, 321
- Seager, S. 2011, *ConPh*, **52**, 602
- Skilling, J. 2004, in *AIP Conf. Proc.*, Vol. 735, Bayesian Inference and Maximum Entropy Methods in Science and Engineering, ed. R. Fischer, R. Preuss, & U. von Toussaint (New York: AIP), 395
- Skilling, J. 2006, *BayAn*, **1**, 833
- Snellen, I. A. G., de Kok, R. J., de Mooij, E. J. W., & Albrecht, S. 2010, *Natur*, **465**, 1049
- Southworth, J. 2010, *MNRAS*, **408**, 1689
- Swain, M. R., Line, M. R., & Deroo, P. 2014, *ApJ*, **784**, 133
- Swain, M. R., Vasish, G., & Tinetti, G. 2008, *Natur*, **452**, 329
- Swain, M. R., Deroo, P., Griffith, C. A., et al. 2010, *Natur*, **463**, 637
- Tennyson, J., & Yurchenko, S. N. 2012, *MNRAS*, **425**, 21
- Ter Braak, C. J. F., & Vrugt, J. A. 2008, *Statistics and Computing*, **18**, 435
- Terriile, R. J., Lee, S., Tinetti, G., et al. 2008, in *2008 IEEE Aerospace Conf.*, 1–9
- Thatte, A., Deroo, P., & Swain, M. R. 2010, *A&A*, **523**, 35
- Tinetti, G., Encrenaz, T., & Coustenis, A. 2013, *A&ARv*, **21**, 63
- Tinetti, G., Tennyson, J., Griffith, C. A., & Waldmann, I. P. 2012, *Philosophical Transactions A*, **370**, 2749
- Turk, M. A., & Pentland, A. P. 1991, in *Proc. CVPR '91*, IEEE Computer Society Conf., Computer Vision and Pattern Recognition, 586
- Venot, O., Agández, M., Selsis, F., Tessenyi, M., & Iro, N. 2014, *A&A*, **562**, 51
- Verde, L., Feeney, S. M., Mortlock, D. J., & Peiris, H. V. 2013, *J. Cosmol. Astropart. Phys.*, **09**, 013
- Verdinelli, I., & Wasserman, L. 2012, *Journal of the American Statistical Association*, **90**, 614
- Waldmann, I. P. 2012, *ApJ*, **747**, 12
- Waldmann, I. P. 2014, *ApJ*, **780**, 23
- Waldmann, I. P., Tinetti, G., Deroo, P., et al. 2013, *ApJ*, **766**, 7
- Waldmann, I. P., Tinetti, G., Drossart, P., et al. 2012, *ApJ*, **744**, 35
- Wark, D. Q., & Hilleary, D. T. 1969, *Sci*, **165**, 1256
- Weinberg, M. D. 2012, *BayAn*, **7**, 737
- Yadin, B., Veness, T., Conti, P., et al. 2012, *MNRAS*, **425**, 34
- Yurchenko, S. N., Barber, R. J., & Tennyson, J. 2011, *MNRAS*, **413**, 1828
- Yurchenko, S. N., & Seager, S. 2014, arXiv:1401.4852v1
- Zhu, C., Byrd, R. H., Lu, P., & Nocedal, J. 1997, *ACM Transactions on Mathematical Software*, Vol. 23 (New York: ACM), 550