

# Tensor-Train Paper

Lennart

November 2025

## 1 Introduction

- Hierarchical Bayesian modelling including Pressure and Temperature
- Forward model ozone limb-sounding as linear inverse problem
- Marginal and full conditional posterior explicitly formulated
- samples from full conditional posterior via IRT
- sample-based affine map
- same procedure again
- full posterior mean and sample-based variance do not capture second ozone peak
- Pressure and Ozone highly correlated
- All programming and analysis in this thesis are done in Python, and the reported computation times are taken on a MacBook Pro from 2019 with a 2.4 GHz quad-core Intel i5 processor.

## 2 Linear inverse Problem – Hierarchical Bayesian Modelling

First the concept of hierarchical Bayesian modelling is introduced. Assume we observe some data

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\eta}, \quad (1)$$

based on a linear forward model  $\mathbf{A}$  an unknown parameter vector  $\mathbf{x}$  and some additive random noise  $\boldsymbol{\eta}$ . Naturally, due to the noise, the observation process in Eq. 1 is a random process. Hence, in Bayesian modelling, the aim is to determine a probability distribution over the parameter  $\mathbf{x}$  given some data  $\mathbf{y}$ . Further, a hierarchical Bayesian model incorporates (auxiliary) hyper-parameters

$\theta$ . Within a Bayesian approach all unknown hyper-parameters and parameters are treated as random variables [10, Chapter 3].

According to Bayes’ theorem, the joint posterior distribution over the parameters  $\mathbf{x}$  and the hyper-parameter  $\theta$  is given as

$$\pi(\mathbf{x}, \theta | \mathbf{y}) = \frac{\pi(\mathbf{y} | \mathbf{x}, \theta) \pi(\mathbf{x}, \theta)}{\pi(\mathbf{y})} \propto \pi(\mathbf{y} | \mathbf{x}, \theta) \pi(\mathbf{x}, \theta), \quad (2)$$

with finite and non-zero  $\pi(\mathbf{y})$ . The likelihood function  $\pi(\mathbf{y} | \mathbf{x}, \theta)$  is defined by the nature of the noise and the noise-free data  $\mathbf{Ax}$ , which we read as the distribution over  $\mathbf{y}$  conditioned on  $\mathbf{x}$  and  $\theta$ . Here  $\theta$  describe multiple hyper-parameters, e.g. the noise precision so that  $\boldsymbol{\eta} \sim \pi_{\boldsymbol{\eta}}(\cdot | \theta)$ , where  $\sim$  reads as “is distributed as”. Further,  $\theta$  may account for some physical properties of  $\mathbf{x}$  such as the smoothness (see Sec. 4). Because all unknown parameter are treated as random variables the joint prior distribution is introduced as  $\pi(\mathbf{x}, \theta) = \pi(\mathbf{x} | \theta) \pi(\theta)$  with the parameter prior distribution  $\pi(\mathbf{x} | \theta)$  and the hyper-prior distribution  $\pi(\theta)$ . Choosing these prior distributions is ultimately a modeller’s choice and is crucial, as those shall be as uninformative as possible for regions in hyper-parameter and parameter space where the data is informative. If the data is uninformative, the prior distributions can be informative and may represent a rather restrictive range of (physically) feasible hyper-parameters and parameters.

We can write the hierarchical model as:

$$\mathbf{y} | \mathbf{x}, \theta \sim \pi(\mathbf{y} | \mathbf{x}, \theta) \quad (3a)$$

$$\mathbf{x} | \theta \sim \pi(\mathbf{x} | \theta) \quad (3b)$$

$$\theta \sim \pi(\theta). \quad (3c)$$

Usually, the objective is to calculate the expectation of a function  $h(\mathbf{x})$ , which is defined as

$$\mathbb{E}_{\mathbf{x}, \theta | \mathbf{y}}[h(\mathbf{x})] = \int \int h(\mathbf{x}) \pi(\mathbf{x}, \theta | \mathbf{y}) \, d\mathbf{x} \, d\theta. \quad (4)$$

## 2.1 Marginal and then Conditional Method

Characterising the posterior distribution or quickly generating a representative sample set from the posterior distribution often presents a significant challenge. This is mainly due to the strong correlations that usually exist between the parameters  $\mathbf{x}$  and hyper-parameters  $\theta$ , as discussed by Rue and Held in [16].

Depending on the problem and the available model it is beneficial to factorise the joint posterior distribution

$$\pi(\mathbf{x}, \theta | \mathbf{y}) = \pi(\mathbf{x} | \theta, \mathbf{y}) \pi(\theta | \mathbf{y}) \quad (5)$$

into the full conditional posterior  $\pi(\mathbf{x} | \theta, \mathbf{y})$  over the latent field  $\mathbf{x}$  and the marginal posterior  $\pi(\theta | \mathbf{y})$  over hyper-parameter  $\theta$ . This approach, known as the marginal and then conditional (MTC) method [6], is particularly advantageous

when  $\mathbf{x} \in \mathbb{R}^n$  is high-dimensional, while  $\boldsymbol{\theta}$  is low-dimensional and the evaluation of the marginal posterior

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})\pi(\mathbf{y})} \propto \frac{\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})} \quad (6)$$

as in [6, Lemma 2] is relatively cheap.

Applying the law of total expectation [5], Eq. (4) becomes

$$\mathbb{E}_{\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}}[h(\mathbf{x})] = \int \int h(\mathbf{x})\pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y}) \, d\mathbf{x} \, \pi(\boldsymbol{\theta}|\mathbf{y}) \, d\boldsymbol{\theta} \quad (7)$$

$$= \int \mathbb{E}_{\mathbf{x}|\boldsymbol{\theta}, \mathbf{y}}[h(\mathbf{x})] \, \pi(\boldsymbol{\theta}|\mathbf{y}) \, d\boldsymbol{\theta} \quad (8)$$

$$= \mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}[\mathbb{E}_{\mathbf{x}|\boldsymbol{\theta}, \mathbf{y}}[h(\mathbf{x})]] \, . \quad (9)$$

In the case of a linear-Gaussian hierarchical Bayesian model, both the marginal distribution  $\pi(\boldsymbol{\theta}|\mathbf{y})$  and the inner expectation  $\mathbb{E}_{\mathbf{x}|\boldsymbol{\theta}, \mathbf{y}}[h(\mathbf{x})]$  are well defined (see Sec. 4). If the integral in Eq. 8 is expensive to calculate sample-based methods may be used to calculate the expectations in Eq. (8). To produce a samples  $\{(\mathbf{x}, \boldsymbol{\theta})^{(1)}, \dots, (\mathbf{x}, \boldsymbol{\theta})^{(k)}, \dots, (\mathbf{x}, \boldsymbol{\theta})^{(N)}\} \sim \pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})$  one needs an independent an sample from  $\boldsymbol{\theta}^{(k)} \sim \pi(\boldsymbol{\theta}|\mathbf{y})$  first and then draw a sample from the full conditional posterior  $\mathbf{x}^{(k)} \sim \pi(\mathbf{x}|\boldsymbol{\theta}^{(k)}, \mathbf{y})$ .

### 3 The Forward Model

Here we present the forward model to which we apply the methodology. The forward model describes a Limb-sounder measuring thermal radiation of ozone to determine the atmospheric ozone concentration. We follow the MIPAS handbook [14] and simulate data according to an idealised cloud-free atmosphere in local thermodynamic equilibrium, assuming a measurement instrument with infinite spectral resolution and no pointing errors. This is a simplified forward model. No other instrument-specific details such as sensor area or antenna response are included because they are not available to us.

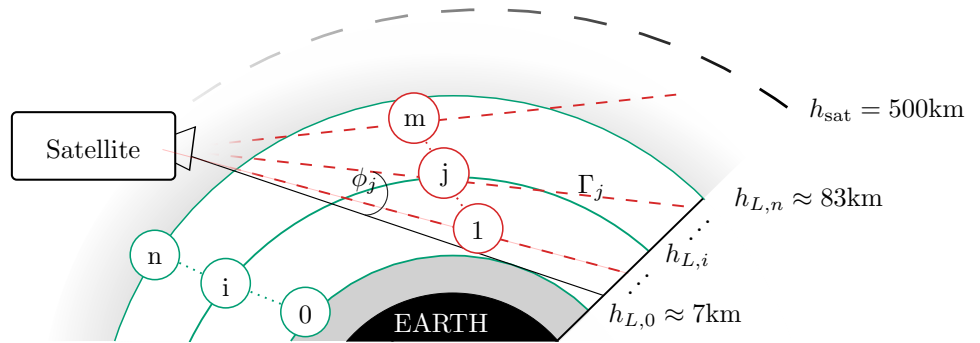


Figure 1: Schematic of measurement and analysis geometry, not to scale. The stationary satellite, at a constant height  $h_{\text{sat}}$  above Earth, takes  $m$  measurements along its line-of-sight defined by the line  $\Gamma_j$ . Each measurement has a pointing angle  $\phi_j$  and a tangent height  $h_{\ell_j}$ ,  $j = 1, 2, \dots, m$  defined as the closest distance of  $\Gamma_j$  to the Earth's surface. Between  $h_{L,0} \approx 7\text{km}$  and  $h_{L,n} \approx 83\text{km}$ , the atmosphere is discretised into  $n$  layers as illustrated by the solid green lines.

As displayed in Fig. 1, a satellite at a constant height  $h_{\text{sat}}$  is pointing through the atmosphere (limb-sounding) to measure thermal radiation of ozone. For each measurement  $j = 1, 2, \dots, m$ , the tangent height  $h_{\ell_j}$  and the corresponding line-of-sight  $\Gamma_j$  are defined. Additionally, we introduce the pointing angle  $0 \leq \phi_j < \phi_{\text{max}}$ , so that if  $\phi = 0$  arc sec the satellite points at  $h_{L,0}$  and for a pointing angle  $\phi_{\text{max}}$  at  $h_{L,n}$ . Further, the atmosphere is discretised into  $n$  layers defined by height values  $h_{L,i-1} < h_{L,i}$  with respect to the surface of the Earth, for  $i = 1, \dots, n$ . More specifically, the  $i$ -th layer is defined by two spheres around the centre of the Earth with radii  $r_0 + h_{L,i-1}$  and  $r_0 + h_{L,i}$ , where  $r_0$  is the Earth's radius. Within a layer the signal is constant, whereas above  $h_{L,n}$  and below  $h_{L,0}$  no signal can be obtained.

### 3.1 Radiative Transfer Equation

One noise-free measurement of thermal radiation emitted by gas molecules within the atmosphere is described by the radiative transfer equation (RTE) [14]

$$\int_{\Gamma_j} B(\nu, T) k(\nu, T) \frac{p(r)}{k_B T(r)} x(r) \tau(r) dr \quad (10)$$

$$\text{with } \tau(r) = \exp \left\{ - \int_{r_{\text{obs}}}^r k(\nu, T) \frac{p(r')}{k_B T(r')} x(r') dr' \right\}. \quad (11)$$

This is a path integral along the satellite's straight line of sight  $\Gamma_j$  with the ozone volume mixing ratio (VMR)  $x(r)$  at distance  $r$  from the satellite, at the wave number  $\nu$ . Within the atmosphere, the number density  $p(r)/(k_B T(r))$  of molecules is dependent on the pressure  $p(r)$ , the temperature  $T(r)$ , and the Boltzmann constant  $k_B$ . The factor  $\tau(r) \leq 1$  accounts for re-absorption of the radiation along the line-of-sight, which makes the RTE non-linear. The absorption constant is given as

$$k(\nu, T) = L(\nu, T_{\text{ref}}) \frac{Q(T_{\text{ref}})}{Q(T)} \frac{\exp \{-c_2 E''/T\}}{\exp \{-c_2 E''/T_{\text{ref}}\}} \frac{1 - \exp \{-c_2 \nu/T\}}{1 - \exp \{-c_2 \nu/T_{\text{ref}}\}} \quad (12)$$

with Planck's constant  $h$  and speed of light  $c$ . The line intensity  $L(\nu, T_{\text{ref}})$  at reference temperature  $T_{\text{ref}} = 296K$ , the lower-state energy  $E''$  in  $\text{cm}^{-1}$  of the targeted transition and the second radiation constant  $c_2 := hc/k_B \approx 1.44\text{cmK}$  are provided by the HITRAN database [8]. The total internal partition function is given as

$$Q(T) = g' \exp \left\{ -\frac{c_2 E'}{T} \right\} + g'' \exp \left\{ -\frac{c_2 E''}{T} \right\}, \quad (13)$$

with the statistical weight  $g''$  for the lower and  $g'$  for the upper energy state (also called the degeneracy factors) accounting for the molecule's non-rotational and rotational energy states (see also [19]), and the upper state energy  $E' = E'' + \nu$ . Under the assumption of local thermodynamic equilibrium (LTE), the black body radiation acts as a source function

$$B(\nu, T) = \frac{2hc^2\nu^3}{\exp \left\{ \frac{c_2 \nu}{T} \right\} - 1}. \quad (14)$$

For fundamentals on the RTE, we recommend [17, Chapter 1], and for a more comprehensive model, we refer to [13].

When simulating data, we assume an idealised limb-sounder. Since the measurement device has a negligible frequency window, the line broadening around  $\nu$  for the calculations of  $L(\nu, T_{\text{ref}})$  is neglected. Normally, this is modelled as the convolution of the normalised Lorentz profile (collisional/pressure broadening) and the normalised Doppler profile (thermal broadening) [14]. Additionally, we target one specific molecule and calculate  $k(\nu, T)$  accordingly. Usually, this would involve a summation over the individual absorption constants for multiple radiating molecules weighted by their respective VMR [14].

### 3.2 Simulated Data and Ground Truth

As the ground truth for our methodology, we consider an ozone profile at distinct pressure values generated from some data [18] of the MLS on the Aura satellite within the Antarctic region. This ozone profile has a peak in the middle atmosphere and a second peak at higher altitudes, see Fig. 15, which seems to be a typical nighttime profile [11]. For more information on the processes within the atmosphere for ozone, we refer to [11].

We can relate the height  $h$  and the pressure values  $p$  via the hydrostatic equilibrium equation

$$d(\log p) = \frac{dp}{p} = \frac{-gM}{R^*T}dh. \quad (15)$$

Here the acceleration due to gravity is  $g$ , the universal gas constant is  $R^* = 8.31432 \times 10^{-3} \text{Nm/kmol/K}$  and the mean molecular weight of the air is  $M = 28.97 \text{kg/kmol}$ , as in [21]. To enable efficient calculation of the RTE we discretise the atmosphere as in Fig. 1. Then the ozone VMR  $\mathbf{x} = \{x_1, x_2, \dots, x_n\} \in \mathbb{R}^n$ , pressure  $\mathbf{p} = \{p_1, p_2, \dots, p_n\} \in \mathbb{R}^n$  and temperature  $\mathbf{T} = \{T_1, T_2, \dots, T_n\} \in \mathbb{R}^n$ , as well as all other height dependent parameters, are discretised profiles with constant values between the heights  $h_{L,i-1} \leq h < h_{L,i}$ , for each layer  $i = 1, \dots, n$ . The hydrostatic equilibrium equation for the discretised atmosphere is

$$h_{L,i} = h_{L,i-1} - \frac{\Delta p R^* T_{i-1}}{p_{i-1} g_{i-1} M} \quad (16)$$

with  $\Delta p = p_i - p_{i-1}$  and  $T_{i-1} = T(h_{i-1})$  as in Eq. 18 (see also [4, 15]), for  $i = 1, \dots, n$ . At sea level  $h = 0 \text{km}$  the mean pressure is  $p_0 = 1013.25 \text{hPa}$  and the mean temperature is  $T_0 = 288.15 \text{K}$  [21]. The acceleration due to gravity is

$$g_i = g_0 \left( \frac{r_0}{r_0 + h_{L,i}} \right), \quad (17)$$

where the polar radius of the Earth is  $r_0 \approx 6356 \text{km}$ , the gravitation at sea level is  $g_0 \approx 9.81 \text{m/s}^2$ . For a ground truth temperature profile we follow [21] and

subscript $i$	geometric height $h_{T,i}$ in km	gradient $a_i$
0	0	-6.5
1	11	0
2	20.1	1
3	32.2	2.8
4	47.4	0
5	51.4	-2.8
6	71.8	-2

Table 1: Definition of height depending temperature gradients.

form the temperature function

$$T(h) = \begin{cases} T_0 & , h = 0 \\ T_0 + a_0 h & , 0 \leq h < h_{T,1} \\ T_0 + a_0 h_{T,1} & , h_{T,1} \leq h < h_{T,2} \\ T_0 + a_0 h_{T,1} + a_1(h_{T,2} - h_{T,1}) + a_2(h - h_{T,2}) & , h_{T,2} \leq h < h_{T,3} \\ T_0 + a_0 h_{T,1} + a_1(h_{T,2} - h_{T,1}) \\ \quad + a_2(h_{T,3} - h_{T,2}) + a_3(h - h_{T,3}) & , h_{T,3} \leq h < h_{T,4} \\ T_0 + a_0 h_{T,1} + a_1(h_{T,2} - h_{T,1}) \\ \quad + a_2(h_{T,3} - h_{T,2}) + a_3(h_{T,4} - h_{T,3}) + a_4(h - h_{T,4}) & , h_{T,4} \leq h < h_{T,5} \\ T_0 + a_0 h_{T,1} + a_1(h_{T,2} - h_{T,1}) \\ \quad + a_2(h_{T,3} - h_{T,2}) + a_3(h_{T,4} - h_{T,3}) + a_4(h_{T,5} - h_{T,4}) \\ \quad + a_5(h - h_{T,5}) & , h_{T,5} \leq h < h_{T,6} \\ T_0 + a_0 h_{T,1} + a_1(h_{T,2} - h_{T,1}) \\ \quad + a_2(h_{T,3} - h_{T,2}) + a_3(h_{T,4} - h_{T,3}) + a_4(h_{T,5} - h_{T,4}) \\ \quad + a_5(h_{T,6} - h_{T,5}) + a_6(h - h_{T,6}) & , h_{T,6} \leq h \lesssim 86 \end{cases} \quad (18)$$

with gradient and height values in Tab. 1 provided by [21]. This function describes the mean temperature in the atmosphere with various height-depending gradients according to the different atmospheric layers. This holds up to a geometric height of 86km, where we ignore a 0.04% non-linear change in  $M$  from 80km to 86km.

We target ozone at a frequency of 235.71GHz, which lies within the region where the MLS observes ozone [12, 23]. The corresponding wave number is  $\nu = 7.86\text{cm}^{-1}$ . The absorption constant  $k(\nu, T)$  is calculated as in Eq. 11, following the high-resolution transmission (HITRAN) database [8]. The HITRAN database provides the line intensity  $L(\nu, T_{\text{ref}})$  for the isotopologue  $^{16}\text{O}_3$  with the AFGL Code 666.

To compute a data vector, we define an atmosphere between  $h_{L,0} = 6.9\text{km}$  and  $h_{L,n} = 83.3\text{km}$  with  $n = 45$  equidistant layers and a satellite fixed at a height of  $h_{\text{sat}} = 500\text{km}$  (see Fig. 1). We measure  $m = 30$  times between heights of  $\approx 7\text{km}$  and  $\approx 68\text{km}$  with pointing accuracy 175arc sec and equidistant spaced

pointing angles

$$\phi_j = (j - 1)175\text{arc sec}, \quad \text{for } j = 1, \dots, 30.$$

Above  $\approx 68\text{km}$  the data is noise dominated (see Fig. 6), hence no measurement are taken in larger altitudes. Each pointing angle  $\phi_j$  defines a path  $\Gamma_j$  (see Fig. 1). The corresponding path integrals in Eq. 10 and Eq. 11 are evaluated using the trapezoidal rule and define the non-linear forward model  $\mathbf{A}(\mathbf{x}, \mathbf{p}, \mathbf{T}) \in \mathbb{R}^m$  for the set of  $m$  noise-free measurements. Here, each entry  $A_j$  of  $\mathbf{A}(\mathbf{x}, \mathbf{p}, \mathbf{T}) \in \mathbb{R}^m$  includes multiple evaluations of the integral in Eq. 11 to calculate the absorption  $\tau(r)$ . For brevity we denote the non-linear forward model as  $\mathbf{A}(\mathbf{x}) := \mathbf{A}(\mathbf{x}, \mathbf{p}, \mathbf{T})$ . The simulated data vector

$$\mathbf{y} = \mathbf{A}(\mathbf{x}) + \boldsymbol{\eta} \quad (19)$$

includes an additive identically-distributed Gaussian noise vector  $\boldsymbol{\eta} \sim \mathcal{N}(0, \gamma^{-1}\mathbf{I})$ . The noise precision is chosen so that the signal-to-noise ratio (SNR) is approximately 150. The SNR is defined as

$$\text{SNR} := \frac{\max(y)}{\text{STD noise}} = \frac{\text{peak signal}}{\text{RMS noise}}, \quad (20)$$

where STD noise is the standard deviation of the noise. An SNR of 150 is similar to [7], where a signal with a maximal spectral intensity of around 100K and a noise range of 0.4 to 1.6K is reported.

By neglecting the absorption (e.g., set  $\tau = 1$  in Eq. (11)) the RTE is linearised. This denotes the linear forward model matrix  $\mathbf{A}_L \in \mathbb{R}^{m \times n}$ . The integral in Eq. (10) is evaluated using the trapezoidal rule and enables matrix-vector multiplication  $\mathbf{A}_L \mathbf{x}$  to compute noise-free linear data. Since neglecting the absorption changes the measurements only slightly (about 1%, see Sec. 6.2), we classify the inverse problem as a weakly non-linear inverse problem. Note that the methods used here will work with different SNRs or other frequencies.



## 4 Hierarchical Bayesian Model

Here a hierarchical Bayesian model is developed where the forward model is described by either  $\mathbf{A}_L$  the linearised RTE (see Eq. 10) or by the approximation  $\mathbf{M}\mathbf{A}_L$  to the non-linear RTE, where  $\mathbf{M}$  is an affine map. Parameters and hyper-parameter are classified and a choice of prior distributions is established. A directed acyclic graph (DAG) is used to visualise conditional dependencies between hyper-parameters and the parameter (see Fig. ??), as well as how distributions progress through to an observation (square box). We plot statistical dependencies as solid arrows and deterministic dependencies as dotted arrows. Then, applying the MTC scheme, we explicitly formulate the respective posterior distributions.

### 4.1 Prior Modelling

First we describe the ozone parameter through normally distributed prior  $\mathbf{x}|\delta \sim \mathcal{N}(0, \mathbf{Q}^{-1})$  with zero mean and no other restrictions, it is clear that our model does not take into account that ozone values cannot be negative. The precision matrix of that prior distribution is

$$\mathbf{Q} = \delta \mathbf{L} = \delta \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix} \quad (21)$$

which is a discrete approximation to the second derivative operator with Dirichlet boundary condition and defines a 1-dimensional Graph Laplacian as in [22, 6], accounting for smoothness in the ozone profile. We reduce the dimension of  $\mathbf{x}$  from 45 to  $n = 34$  by discarding every second ozone VMR over a height of  $\approx 47\text{km}$ . Doing that, while not changing  $\mathbf{L}$  effectively induces a larger correlation between points at higher altitude.

Since pressure and temperature are treated as unknowns, they are included within the hierarchical structure and represented through model describing hyper-parameters. We observe that the pressure  $\mathbf{p}$  in between  $h_{L,0} \approx 7\text{km}$  and  $h_{L,n} \approx 83\text{km}$  can be described with an exponential function

$$p(h) = \exp(-bh) p_0, \quad h_{L,0} \leq h \leq h_{L,n} \quad (22)$$

depending on two hyper-parameters  $p_0, b$  (see Fig. 3). Similarly, the temperature as described in Eq. 18 can be parametrised with 14 hyper-parameters  $\mathbf{h}_T = \{h_{T,1}, h_{T,2}, h_{T,3}, h_{T,4}, h_{T,5}, h_{T,6}\}$ ,  $\mathbf{a} = \{a_0, a_1, a_2, a_3, a_4, a_5, a_6\}$  and  $T_0$  (see Fig. 2 and Eq. 18).

The hyper-prior distributions for  $p_0, b, T_0, \mathbf{h}_T, \mathbf{a}$  are defined to be Gaussians, and to complete the model we have to choose sensible hyper-prior variances and means. The variances of  $\pi(\mathbf{h}_T)$  are tuned so that the temperature profile maintains its structure and  $h_{T,i} < h_{T,i+1}$ , for  $i = 1, \dots, 5$  (see Fig. ??). The

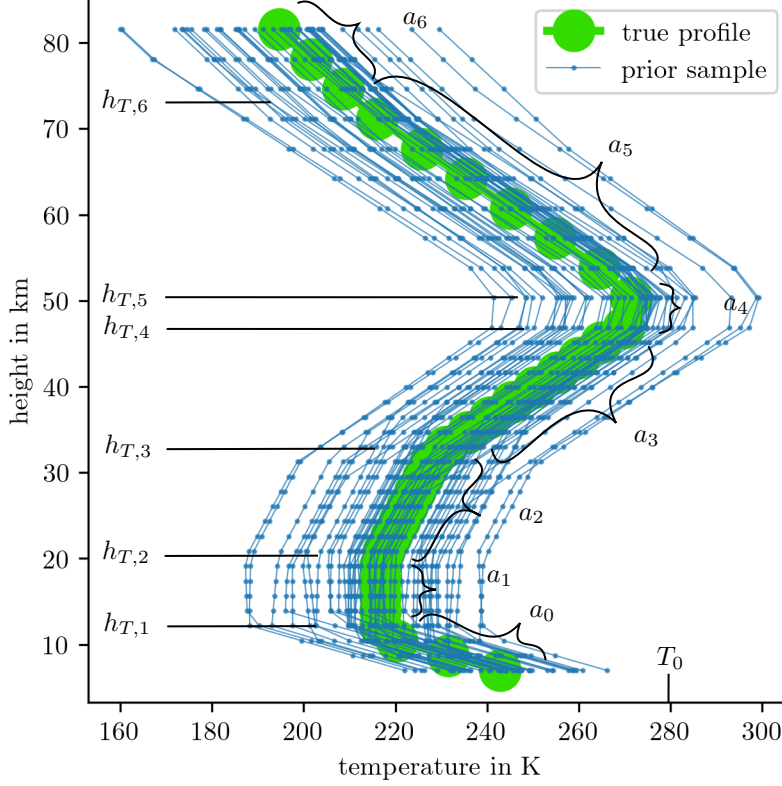


Figure 2: Prior samples from the hyper-prior distribution of  $\mathbf{h}_T$ ,  $\mathbf{a}$  and  $T_0$ , as defined in Tab. 2, where we calculate  $\mathbf{T}$  according to the function in Eq. 18.

means of  $\pi(\mathbf{h}_T)$  and  $\pi(\mathbf{a})$  are set to ground truth values see Tab. 1 and the variances of  $\pi(\mathbf{a})$  allow a wide range of prior temperature profiles. Similarly, the variance and mean of  $\pi(T_0)$  are chosen to mimic a daily temperature variability of roughly 30K around the mean sea level temperature 288K [21]. These hyper-prior distributions are rather informative, because we find that the data and the model (see Fig. 4) are uninformative about the temperature profile. The variance of  $\pi(b)$  is set to a rather large value. The variability of  $\pi(p_0)$  is  $\approx 80$ hPa and close to what we can observe when looking at weather data. Means for  $\pi(b, p_0)$  are provided by fitting the exponential in Eq. 22 to ground truth pressure values via the Python function `scipy.optimize.curve_fit`.

Prior samples against their ground truth profiles of the pressure  $\mathbf{p}$  are plotted in Fig. 3, the temperature  $\mathbf{T}$  in Fig. 2, the ratio  $\mathbf{p}/\mathbf{T}$  in Fig. 4 and additionally prior samples of  $1/\mathbf{T}$  are plotted in Fig. ?? . In Fig. 4 we already observe that  $\mathbf{p}/\mathbf{T}$  inherits the structure of the pressure function and hence the model is uninformative about the temperature.

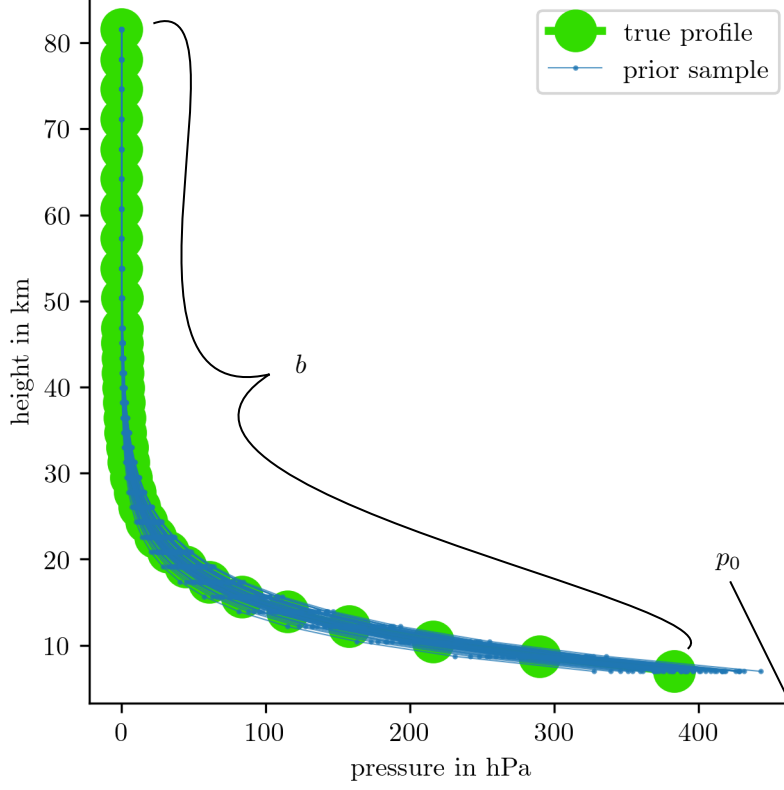


Figure 3: Prior samples from the hyper-prior distribution of  $b$  and  $p_0$  as defined in Tab. 2, where we calculate  $\mathbf{p}$  according to the function in Eq. 22.

For  $\delta$  and  $\gamma$  we pick relatively uninformative Gamma distributions so that  $\gamma \sim \mathcal{T}(\boldsymbol{\theta}_\gamma) \propto \gamma^{\alpha_\gamma - 1} \exp(-\beta_\gamma \gamma)$  and  $\delta \sim \mathcal{T}(\boldsymbol{\theta}_\delta)$  with  $\boldsymbol{\theta}_\gamma = \{\alpha_\gamma, \beta_\gamma\} = \{\alpha_\delta, \beta_\delta\} = \boldsymbol{\theta}_\delta = (1, 10^{-35})$  similar to [6].

See Tab. 2 for a summary of the prior distributions.

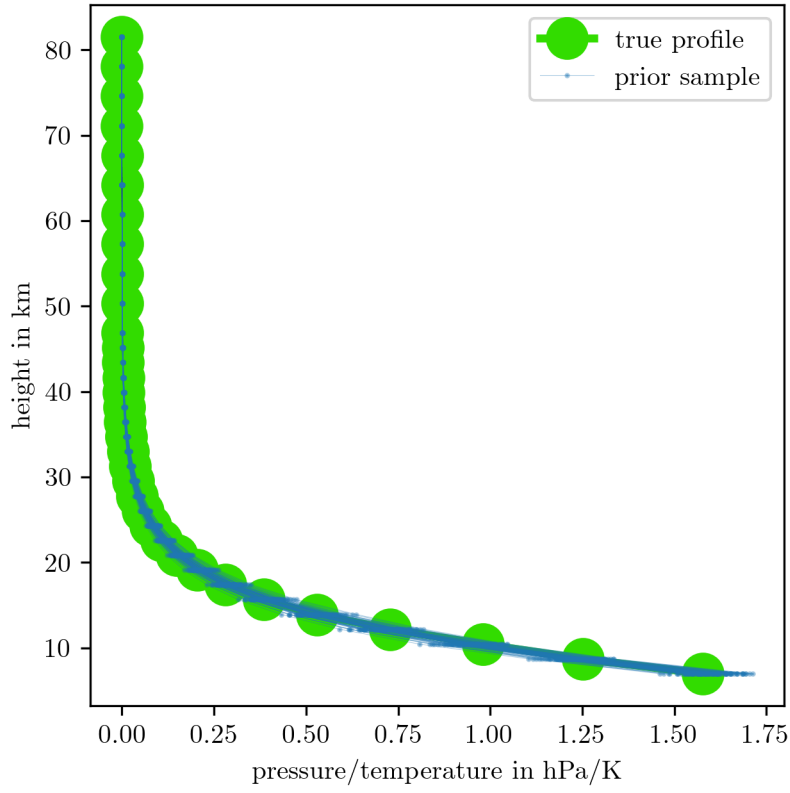


Figure 4: Prior samples from the hyper-prior distribution of  $\mathbf{h}_T$ ,  $\mathbf{a}$  and  $T_0$  for temperature as in Eq. 18 and  $b$  and  $p_0$  for pressure as in Eq. 22. We plot  $\mathbf{p}/\mathbf{T}$ . The hyper-priors are defined in Tab. 2.

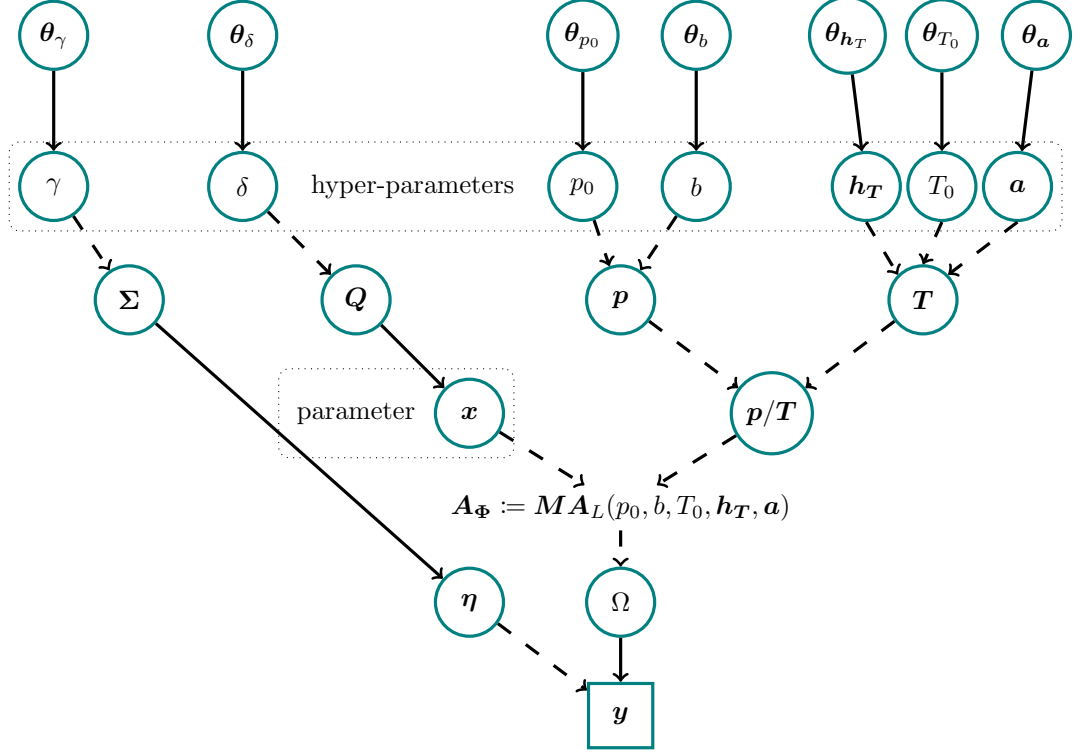


Figure 5: DAG of the hierarchical Bayesian model including ozone  $x$ , pressure  $p$  and temperature  $T$ . The hyper-parameters  $\mathbf{h}_T = \{h_{T,1}, h_{T,2}, h_{T,3}, h_{T,4}, h_{T,5}, h_{T,6}\}$ ,  $\mathbf{a} = \{a_0, a_1, a_2, a_3, a_4, a_5, a_6\}$ ,  $T_0$ ,  $b$  and  $p_0$  deterministically (dotted line) describe pressure through Eq. 22 and temperature through Eq. 18. In this case, the hyper-parameters  $\pi(p_0, b, T_0, \mathbf{h}_T, \mathbf{a})$  are a normally distributed a-priori. That is why  $\theta_{h_T}, \theta_a, \theta_{T_0}, \theta_b, \theta_{p_0}$  represent means and STDs e.g.,  $b \sim \mathcal{N}(\mu_b, \sigma_b^2)$  and  $\theta_b = \{\mu_b, \sigma_b\}$ . As previously described in Sec. ??,  $\theta_\gamma, \theta_\delta$  determine gamma distributions e.g.,  $\gamma \sim \mathcal{T}(\alpha_\gamma, \beta_\gamma)$  with  $\theta_\gamma = \{\alpha_\gamma, \beta_\gamma\}$ . The ozone parameter  $x$  is statistically (solid line) described by the prior distribution  $x|\delta \sim \mathcal{N}(0, (\delta \mathbf{L})^{-1})$ . Here, the hyper-parameter  $\delta$  accounts for smoothness in the ozone profile and defines the precision matrix  $\mathbf{Q} = \delta \mathbf{L}$  as in Eq. 21. The approximated forward model  $\mathbf{A}_\Phi := \mathbf{M} \mathbf{A}_L(p_0, b, T_0, \mathbf{h}_T, \mathbf{a})$  with  $\Phi := \{p_0, b, T_0, \mathbf{h}_T, \mathbf{a}\}$  maps the parameter and hyper-parameters to a space of all measurables  $\Omega$ . From that space, a data set  $\mathbf{y}$  including some additive random noise  $\boldsymbol{\eta}$  is observed (square box). The noise covariance  $\boldsymbol{\Sigma} = \gamma^{-1} \mathbf{I}$  of the random noise vector  $\boldsymbol{\eta} \sim \mathcal{N}(0, \gamma^{-1} \mathbf{I})$  is defined by the hyper-parameter  $\gamma$ .

As in Sec. ??, the DAG in Fig. 5 visualises the measurement process and conditional dependencies between the parameter and the hyper-parameters. This hierarchical Bayesian framework includes the hyper-parameters  $p_0, b$  for pressure (see Eq. 22),  $\mathbf{a}, \mathbf{h}_T, T_0$  for temperature (see Eq. 18),  $\delta$  for ozone smoothness and

$\gamma$  for noise precision. The hyper-parameters are described by the hyper-prior distribution  $\pi(p_0, b, T_0, \mathbf{h}_T, \mathbf{a}, \delta, \gamma)$  (see Sec. 4.1). Through their respective prior distributions, pressure  $\mathbf{p}$ , temperature  $\mathbf{T}$  and ozone  $\mathbf{x}$  progress deterministically (dashed line) into the forward model via  $\mathbf{x} \times \mathbf{p}/\mathbf{T}$  and generate a space of all possible noise-free data  $\Omega$ . Note that other variables in the RTE, such as the internal partition function and the black body radiation, are dependent on temperature (see Sec. 3.1). From  $\Omega$  we observe (square box) some data with additive normally distributed noise  $\boldsymbol{\eta}$ . For brevity, we define the linear forward model matrix as

$$\mathbf{A}_\Phi := \mathbf{M}\mathbf{A}_L(p_0, b, T_0, \mathbf{h}_T, \mathbf{a}) \quad (23)$$

with  $\Phi := \{p_0, b, T_0, \mathbf{h}_T, \mathbf{a}\}$  accounting for the all pressure and temperature related hyper-parameters and  $\mathbf{M}$  is an affine approximation. First  $\mathbf{M} = \mathbf{I}$  is set to the identity matrix and after obtained results with a model given by the linearised RTE updated to an affine approximation mapping from the linearised RTE to the non-linear RTE.

The distributions of the hierarchical Bayesian framework are:

$$\mathbf{y}|\mathbf{x}, \Phi, \delta, \gamma \sim \mathcal{N}(\mathbf{A}_\Phi \mathbf{x}, \gamma^{-1} \mathbf{I}) \quad (24a)$$

$$\mathbf{x}|\delta \sim \mathcal{N}(\mathbf{0}, (\delta \mathbf{L})^{-1}) \quad (24b)$$

$$\delta \sim \mathcal{T}(\alpha_\delta, \beta_\delta) \quad (24c)$$

$$\gamma \sim \mathcal{T}(\alpha_\gamma, \beta_\gamma) \quad (24d)$$

$$\mathbf{a} \sim \mathcal{N}(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a) \quad (24e)$$

$$\mathbf{h}_T \sim \mathcal{N}(\boldsymbol{\mu}_T, \boldsymbol{\Sigma}_{h_T}) \quad (24f)$$

$$T_0 \sim \mathcal{N}(\mu_{T_0}, \sigma_{T_0}^2) \quad (24g)$$

$$p_0 \sim \mathcal{N}(\mu_{p_0}, \sigma_{p_0}^2) \quad (24h)$$

$$b \sim \mathcal{N}(\mu_b, \sigma_b^2). \quad (24i)$$

Due to Gaussian noise  $\pi(\mathbf{y}|\mathbf{x}, \Phi, \delta, \gamma)$  is a normally distributed likelihood function and Eq. 24b to Eq. 24i denote prior distributions. Before formulating the posterior distribution, we carefully define  $\boldsymbol{\theta}_\gamma, \boldsymbol{\theta}_\delta, \boldsymbol{\theta}_{p_0}, \boldsymbol{\theta}_b, \boldsymbol{\theta}_h, \boldsymbol{\theta}_{T_0}, \boldsymbol{\theta}_a$ , the hyper-prior scales, shapes, means and variances, which are explicitly given in Tab. 2.

## 5 Posterior Distribution

Here, the marginal posterior over the hyper-parameters  $\boldsymbol{\theta} = \{\Phi, \gamma, \delta\}$  and the full conditional posterior distribution for the parameter  $\mathbf{x}|\boldsymbol{\theta}$  are formulated. To draw samples from the marginal posterior  $\pi(\boldsymbol{\theta}|\mathbf{y})$  we utilise a TT approximation on a predefined grid to generate samples via the SIRT method with an MH correction step. In doing so, the reader is guided through the process of obtaining an efficient higher-dimensional TT approximation. Lastly, the RTO method is utilised to draw ozone samples from the full conditional posterior  $\pi(\mathbf{x}|p_0, b, T_0, \mathbf{h}_T, \mathbf{a}, \delta, \gamma, \mathbf{y})$ . Recall that  $\mathbf{x} \in \mathbb{R}^n$  with  $n = 34$  and  $\mathbf{y} \in \mathbb{R}^m$

with  $m = 30$ , and that the linear forward model matrix  $\mathbf{A}_{\Phi}$  is depending on the hyper-parameter defined as  $\Phi := \{p_0, b, T_0, \mathbf{h}_T, \mathbf{a}\}$ . Consequently, the marginal posterior is denoted as  $\pi(\Phi, \delta, \gamma | \mathbf{y})$  and the full conditional posterior as  $\pi(\mathbf{x} | \Phi, \delta, \gamma, \mathbf{y})$ .

model parameters	priors	TT bounds		Context
		lower	upper	
$\mathbf{x}$	$\mathcal{N}(0, (\delta \mathbf{L})^{-1})$	-	-	$\mathbf{x}$
$\delta$	$\mathcal{T}(1, 10^{-35})$	-	-	$\mathbf{x}$
$\gamma$	$\mathcal{T}(1, 10^{-35})$	$8 \times 10^{14}$	$1.2 \times 10^{16}$	$\mathbf{y}$
$b$	$\mathcal{N}(0.174, (0.01)^2)$	0.129	0.214	$\mathbf{p}$
$h_{T,1}$	$\mathcal{N}(11, (1.5)^2)$	5.4	16.3	$\mathbf{T}$
$T_0$	$\mathcal{N}(288.15, (10)^2)$	247	326	$\mathbf{T}$
$p_0$	$\mathcal{N}(1311, (20)^2)$	1237	1387	$\mathbf{p}$
$h_{T,3}$	$\mathcal{N}(32.3, (2.5)^2)$	22.9	41.7	$\mathbf{T}$
$a_1$	$\mathcal{N}(0, (0.1)^2)$	-0.38	0.38	$\mathbf{T}$
$h_{T,2}$	$\mathcal{N}(20.1, (0.7)^2)$	17.2	22.7	$\mathbf{T}$
$a_0$	$\mathcal{N}(-6.5, (0.01)^2)$	-6.54	-6.47	$\mathbf{T}$
$a_2$	$\mathcal{N}(1, (0.01)^2)$	0.97	1.03	$\mathbf{T}$
$a_3$	$\mathcal{N}(2.8, (0.1)^2)$	2.5	3.1	$\mathbf{T}$
$h_{T,4}$	$\mathcal{N}(47.4, (0.5)^2)$	45.5	49.3	$\mathbf{T}$
$a_4$	$\mathcal{N}(0, (0.1)^2)$	-0.38	0.38	$\mathbf{T}$
$h_{T,5}$	$\mathcal{N}(51.4, (0.5)^2)$	49.5	53.3	$\mathbf{T}$
$a_5$	$\mathcal{N}(-2.8, (0.1)^2)$	-3.18	-2.43	$\mathbf{T}$
$h_{T,6}$	$\mathcal{N}(71.8, (3)^2)$	60.5	83.1	$\mathbf{T}$
$a_6$	$\mathcal{N}(-2, (0.01)^2)$	-2.04	-1.96	$\mathbf{T}$

Table 2: Summary of relevant parameter and hyper-parameters bounds and statistics, ordered as in the TT format according to their correlation structure. We denote  $\mathcal{N}(\mu = \text{mean}, \sigma^2 = \text{variance})$  as the Gaussian and  $\mathcal{T}(\alpha = \text{scale}, \beta = \text{rate})$  as the Gamma distribution. The IACT  $\tau_{\text{int}}$  from marginal posterior samples via the t-walk is twice the value provided by [UwerrM, 9].

## 5.1 Marginal Posterior – Pressure and Temperature

The marginal posterior is given as

$$\pi(\Phi, \delta, \gamma | \mathbf{y}) \propto \delta^{n/2} \gamma^{m/2} \exp \left\{ -\frac{1}{2} g(\Phi, \delta, \gamma) - \frac{\gamma}{2} f(\Phi, \delta, \gamma) \right\} \pi(\Phi, \delta, \gamma), \quad (25)$$

with,

$$f(\Phi, \delta, \gamma) = \mathbf{y}^T \mathbf{y} - (\mathbf{A}_{\Phi}^T \mathbf{y})^T (\gamma \mathbf{A}_{\Phi}^T \mathbf{A}_{\Phi} + \delta \mathbf{L})^{-1} (\mathbf{A}_{\Phi}^T \mathbf{y}), \quad (26a)$$

$$\text{and } g(\Phi, \delta, \gamma) = \log \det (\gamma \mathbf{A}_{\Phi}^T \mathbf{A}_{\Phi} + \delta \mathbf{L}). \quad (26b)$$

For each evaluation of  $\pi(\Phi, \delta, \gamma | \mathbf{y})$ ,  $\mathbf{A}_{\Phi}$  is composed as in Chapter 3, and  $f$  and  $g$  are calculated directly using the Cholesky decomposition via the Python functions `numpy.linalg.cholesky` and `scipy.linalg.cho_solve`.

## 5.2 Posterior Distribution

As explained in Sec. 2.1, we factorise the posterior

$$\pi(\mathbf{x}, \Phi, \delta, \gamma | \mathbf{y}) \propto \pi(\mathbf{y} | \mathbf{x}, \Phi, \delta, \gamma) \pi(\mathbf{x}, \Phi, \delta, \gamma) \quad (27)$$

into

$$\pi(\mathbf{x}, \Phi, \delta, \gamma | \mathbf{y}) = \pi(\mathbf{x} | \Phi, \delta, \gamma, \mathbf{y}) \pi(\Phi, \delta, \gamma | \mathbf{y}) \quad (28)$$

the marginal posterior  $\pi(\Phi, \delta, \gamma | \mathbf{y})$  and full conditional posterior  $\pi(\mathbf{x} | \Phi, \delta, \gamma, \mathbf{y})$  (see Eq. 5). As discussed in [6], for the linear-Gaussian case,  $\mathbf{x}$  cancels in the marginal posterior over the hyper-parameters. Following the MTC scheme, we characterise the marginal posterior first and then the full conditional posterior.

### 5.2.1 Full Conditional Posterior

As in [20], consider the joint Gaussian distribution

$$\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \sim \mathcal{N} \left[ \begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{A}\boldsymbol{\mu} \end{pmatrix}, \begin{pmatrix} \mathbf{Q}_{\theta} + \mathbf{A}_{\Phi}^T \boldsymbol{\Sigma}_{\theta}^{-1} \mathbf{A}_{\Phi} & -\mathbf{A}_{\Phi}^T \boldsymbol{\Sigma}_{\theta}^{-1} \\ \boldsymbol{\Sigma}_{\theta}^{-1} \mathbf{A}_{\Phi} & \boldsymbol{\Sigma}_{\theta}^{-1} \end{pmatrix}^{-1} \right] \quad (29)$$

with  $\boldsymbol{\Sigma}_{\theta}^{-1} = \gamma \mathbf{I}$  and  $\mathbf{Q}_{\theta} = \delta \mathbf{L}$  and  $\boldsymbol{\mu} = \mathbf{0}$ . Then the full conditional posterior distribution of ozone

$$\mathbf{x} | \Phi, \delta, \gamma, \mathbf{y} \sim \mathcal{N}((\mathbf{A}_{\Phi}^T \mathbf{A} + \delta/\gamma \mathbf{L})^{-1} \mathbf{A}_{\Phi}^T \mathbf{y}, (\gamma \mathbf{A}_{\Phi}^T \mathbf{A}_{\Phi} + \delta \mathbf{L})), \quad (30)$$

is a normal distribution and samples can be drawn via the randomise then optimise (RTO) method (see Sec. 6.1.2).



## 6 Results

In this section the results are presented. Given the data see Sec. the inverse problem is treated as a linear inverse problem by neglecting the non-linear absorption in the RTE. Then an affine approximation is found and the non-linear forward model is approximated by the affine map and the linear forward model. Finally, we present posterior ozone, pressure and temperature profiles based on the approximated forward model.

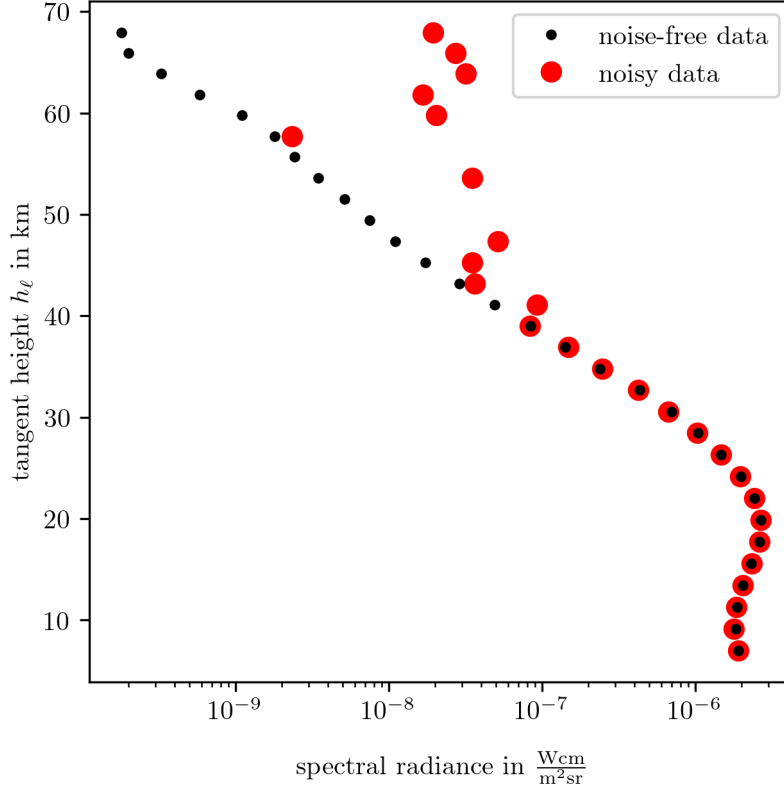


Figure 6: Non-linear noise-free data and non-linear noisy data at different tangent heights. Negative noisy data values are not shown due to logarithmic scaling.

### 6.1 Linear Forward Model

Here as already mentioned the forward model is defined as:

$$\mathbf{A}_\Phi := \mathbf{A}_L(p_0, b, T_0, \mathbf{h}_T, \mathbf{a}). \quad (31)$$

Consequently, with the priors and simulated data, the marginal and conditional posterior are well defined and a Tensor-Train approximation of the marginal posterior is approximated using a tensor-train (TT) and the squared inverse Rosenblatt transform (SIRT) [**<empty citation>**] is employed to generate samples from the marginal posterior. Conditioned on those samples the randomise then optimise (RTO) method [**<empty citation>**] is used to sample from the full conditional posterior.

### 6.1.1 Tensor-Train Approximation of the Marginal Posterior

Instead of using conventional Markov-Chain Monte-Carlo methods to generate samples from the marginal posterior, it is approximated directly using a TT representation. In the following a brief introduction to the TT format is provided.

Assume the target function is  $\pi(\boldsymbol{\Phi}, \delta, \gamma | \mathbf{y})$  and the input  $\boldsymbol{\theta} := \{\boldsymbol{\Phi}, \delta, \gamma\} \in \mathbb{R}^d$  is  $d$ -dimensional. First, a  $d$ -dimensional grid with  $n$  grid points in each dimension the parameter space is defined by  $n^d$  grid points. This is computationally not feasible to represent naively in a  $d$ -dimensional tensor, so a TT is employed to approximate the target function on that parameter space. In practice, a TT approximation may have negative or zero values in regions where the true function value is small and positive. That is why, to ensure non-negativity, we approximate the square root of  $\pi(\boldsymbol{\theta} | \mathbf{y})$  over the whole parameter space, which can be written as:

$$\sqrt{\pi(\boldsymbol{\theta} | \mathbf{y})} \approx \tilde{g}(\boldsymbol{\theta}) = \mathbf{G}_1(\theta_1), \dots, \mathbf{G}_k(\theta_k), \dots, \mathbf{G}_d(\theta_d), \quad (32)$$

where each  $\mathbf{G}_k \in \mathbb{R}^{r_{k-1} \times n \times r_k}$  is a TT core. Further,  $r_k$  are the ranks of the individual cores and bounded by  $r_0 = 1$  and  $r_d = 1$ . For now, assume the ranks  $r_k = r$  for  $k = 1, \dots, d-1$ , then the TT approximates the function space with  $n \times r^2 \times (d-2) + n \times r \times 2$  instead of the  $n^d$  function evaluations.

Given a TT approximation, samples from the target function are generated via the inverse Rosenblatt transform (IRT) scheme [**<empty citation>**], more precisely, the squared inverse Rosenblatt transform (SIRT) since we approximate the square root of the target function. Within this scheme a sequence of 1-dimensional functions can be constructed, so that

$$\begin{aligned} \pi(\boldsymbol{\theta} | \mathbf{y}) = & \pi(\theta_1 | \mathbf{y}) \\ & \times \pi(\theta_2 | \theta_1, \mathbf{y}) \\ & \times \pi(\theta_3 | \theta_2, \theta_1, \mathbf{y}) \\ & \times \pi(\theta_4 | \theta_3, \theta_2, \theta_1, \mathbf{y}) \\ & \vdots \\ & \times \pi(\theta_d | \theta_{d-1}, \dots, \theta_4, \theta_3, \theta_2, \theta_1, \mathbf{y}). \end{aligned} \quad (33)$$

We call each function  $\pi(\theta_k | \theta_{k-1}, \dots, \theta_1, \mathbf{y}) \in \mathbb{R}^n$  the 'conditional marginal' because it is conditioned on  $\theta_{k-1}, \dots, \theta_1$ , and marginalised over  $\theta_{k+1}, \dots, \theta_d$ . To

obtain these is computationally relatively cheap because each of the TT cores is representative for one dimension in the parameter space and hence it is easy to integrate over the individual parameters  $\theta_k \in \boldsymbol{\theta}$  for  $k = 1, \dots, d$ . The marginal distribution of the first dimension is computed as:

$$\pi(\theta_1|\mathbf{y}) \approx \frac{1}{z} G_1^2(\theta_1) \int G_2^2(\theta_2) d\theta_2 \cdots \int G_d^2(\theta_d) d\theta_d, \quad (34)$$

where each of the integrals reduce a core  $\mathbf{G}_k \in \mathbb{R}^{r_{k-1} \times n \times r_k}$  to  $\int \mathbf{G}_k^2(\theta_k) d\theta_k \in \mathbb{R}^{r_{k-1} \times r_k}$  and  $z$  is a normalisation constant see [**<empty citation>**]. If you want to use a quadrature scheme to calculate expectations, e. g.,  $\mathbb{E}_{\pi(\boldsymbol{\theta}|\mathbf{y})}[f(\boldsymbol{\theta})]$ , or are only interested in the individual marginals you are done here since other marginals can be computed the same way. Note that, for a specific parameter value  $\theta_k$  the core  $\mathbf{G}_k^2(\theta_k) \in \mathbb{R}^{r_{k-1} \times r_k}$  is given according to the used interpolation basis, see, e.g., [**<empty citation>**] for piecewise linear interpolation.

The 'conditional marginal' defines a mapping from the uniform distribution to the target function, which is constructed via the cumulative distribution function (CDF)

$$F(\theta_k) = \int_{-\infty}^{\theta_k} \pi(\hat{\theta}_k | \theta_{k-1}, \dots, \theta_1, \mathbf{y}) d\hat{\theta}_k. \quad (35)$$

Then a uniformly distributed random variable  $\mathbf{u} \sim \mathcal{U}[0, 1]^d$  is projected onto the parameter space via the inverse

$$\theta_k = F^{-1}(u_k)$$

with  $u_k \in \mathbf{u}$ . This scheme is repeated for each 'conditional marginal' in Eq. ?? until a sample  $\boldsymbol{\theta} \sim \tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$  is obtained.

Since the samples  $\{\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(N)}\} \sim \tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$  are drawn from an approximation they are corrected with an Metropolis–Hastings (MH) algorithm [**<empty citation>**]. To take one MH correction step from a current state  $\boldsymbol{\theta}$  to a proposed state  $\boldsymbol{\theta}'$  the ratio

$$h(\boldsymbol{\theta}, \boldsymbol{\theta}') = \frac{\pi(\boldsymbol{\theta}'|\mathbf{y})\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})}{\pi(\boldsymbol{\theta}|\mathbf{y})\tilde{\pi}(\boldsymbol{\theta}'|\mathbf{y})} \quad (36)$$

is computed and the proposal  $\boldsymbol{\theta}'$  is accepted with probability

$$h(\boldsymbol{\theta}, \boldsymbol{\theta}') = \min(h(\boldsymbol{\theta}, \boldsymbol{\theta}'), 1). \quad (37)$$

If  $\boldsymbol{\theta}'$  is accepted the next current state is  $\boldsymbol{\theta} = \boldsymbol{\theta}'$ , if  $\boldsymbol{\theta}'$  is rejected  $\boldsymbol{\theta}$  remains the current state.

In practise a small constant  $\xi$  according to the L2-norm error of the TT approximation is added so that

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \approx \tilde{\pi}(\boldsymbol{\theta}|\mathbf{y}) = \frac{1}{z}(\xi + \tilde{g}(\boldsymbol{\theta})^2). \quad (38)$$

This has the consequence that the chain of the MH correction algorithm is uniform geometric convergent (and ergodic [[empty citation](#)]) towards  $\pi(\boldsymbol{\theta}|\mathbf{y})$ , because

$$\tilde{w} = \left\| \frac{\pi(\boldsymbol{\theta}|\mathbf{y})}{\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})} \right\|_{\infty} = \sup \frac{\pi(\boldsymbol{\theta}|\mathbf{y})}{\xi + \tilde{g}(\boldsymbol{\theta})^2} < \infty \quad (39)$$

is bounded [[empty citation](#)]. Without  $\xi$  the TT approximation  $\tilde{g}(\boldsymbol{\theta})^2$  may be arbitrarily small in regions where the true function is small and the ratio in ?? can not be bounded. With  $\xi$  we introduce a more or less uniform error. Further [[empty citation](#)] show that if  $\xi$  is chosen according to the 2-norm error

$$\left\| \tilde{g}(\boldsymbol{\theta}) - \sqrt{\pi(\boldsymbol{\theta}|\mathbf{y})} \right\|_2 \leq \epsilon \quad (40)$$

then the approximation satisfies the error bound

$$\left\| \sqrt{\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})} - \sqrt{\pi(\boldsymbol{\theta}|\mathbf{y})} \right\|_2 \leq \sqrt{2}\epsilon. \quad (41)$$

To generate a TT for  $\sqrt{\pi(\boldsymbol{\theta}|\mathbf{y})}$  we use the [toolbox](#), which implements the TT-Cross algorithm from [[empty citation](#)]. As already mentioned we use the [TT toolbox](#), which uses the TT-Cross amen scheme to approximate the target function. By exploratory analysis we define ranks  $r =$ , grid size  $n =$  and do one sweep (left-to-right and right-to-left) in the TT-Cross. On a [Laptop xxxx](#) function evaluate takes about 1min. To sample from the TT approximation we generate 2000 samples from with the SIRT scheme and correct those with an importance sampling step, as in [[empty citation](#)].

In theory, this scheme provides independent samples, but in practice, every system produces correlated samples. To assess how efficient this scheme is, we define the Integrated Autocorrelation Times (IACT)

Further we calculate the integrated autocorrelation time (IACT) as in [[empty citation](#)] using the code from [[empty citation](#)]. Note that we define the IACT

$$\tau_{\text{int}} := \left( 1 + 2 \sum_{t=1}^W \frac{\Gamma(t)}{\Gamma(0)} \right) \quad (42)$$

as in [6] with an autocorrelation coefficient  $\Gamma(t)$ . This is twice the value of the IACT in [25, pp. 103-105] and [24, 9], as commonly defined within the physics community. U. Wolff [24] (and the Python implementation by D. Hesse [9]) provide a way to not only calculate the IACT safely but also to quantify the errors of the estimated IACT. The IATCs are  $\tau_{\text{int},\lambda} = 1.05 \pm 0.04$  and  $\tau_{\text{int},\gamma} = 0.99 \pm 0.03$  based on a chain with 10000 samples. So every second sample via the IRT scheme presents an independent marginal posterior sample. Conditioned on an independent sample from the marginal posterior, one can draw an independent sample from the full conditional posterior.

Further we can use that to bound the maximum absolute error  $\epsilon$  as in.

$$\frac{\tau_{\text{int}} - 1}{2} \leq \tilde{w} - 1 \quad (43)$$

We can use

$$\tau_{\text{int}} - 1 \quad (44)$$

as an error estimate.

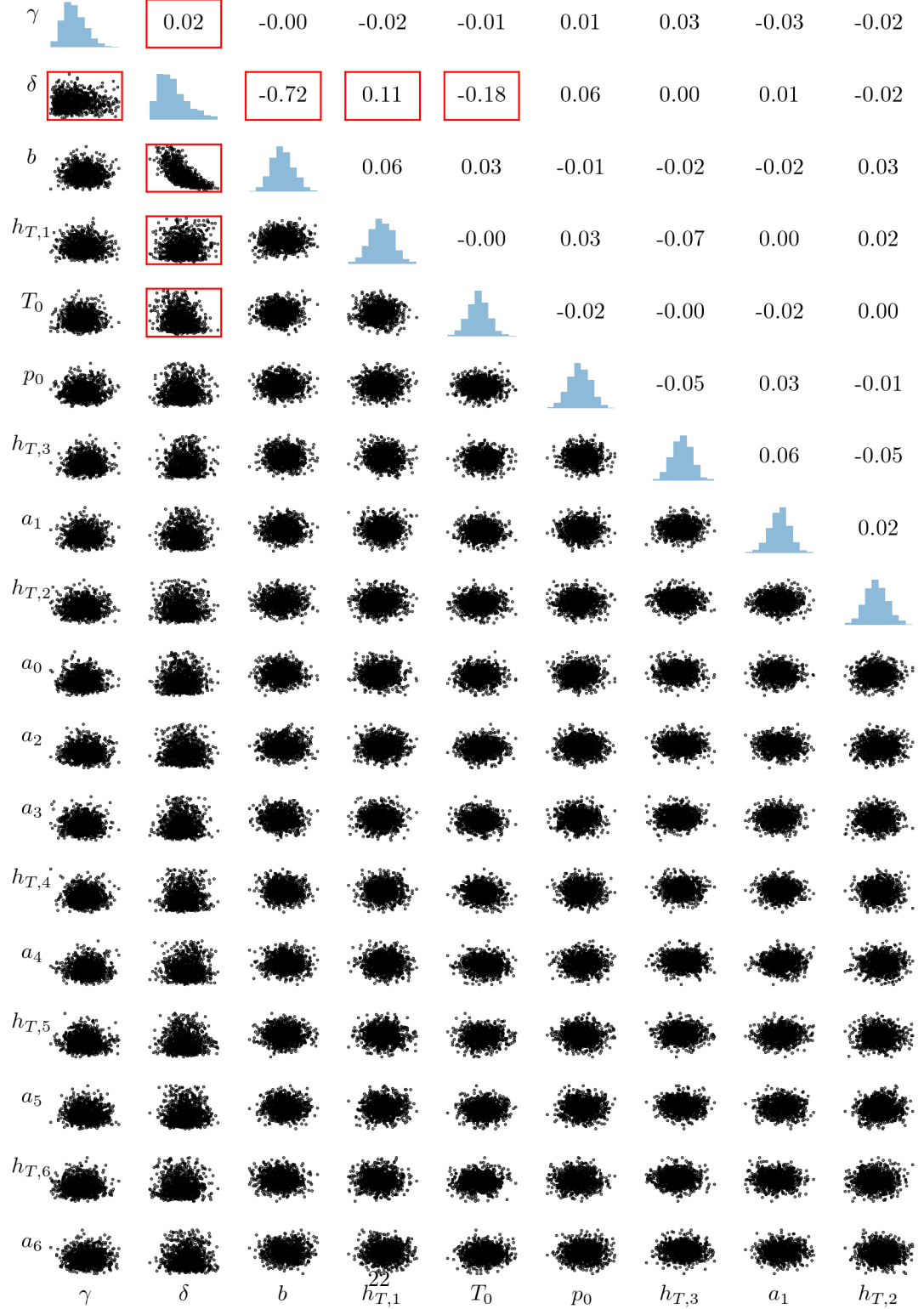
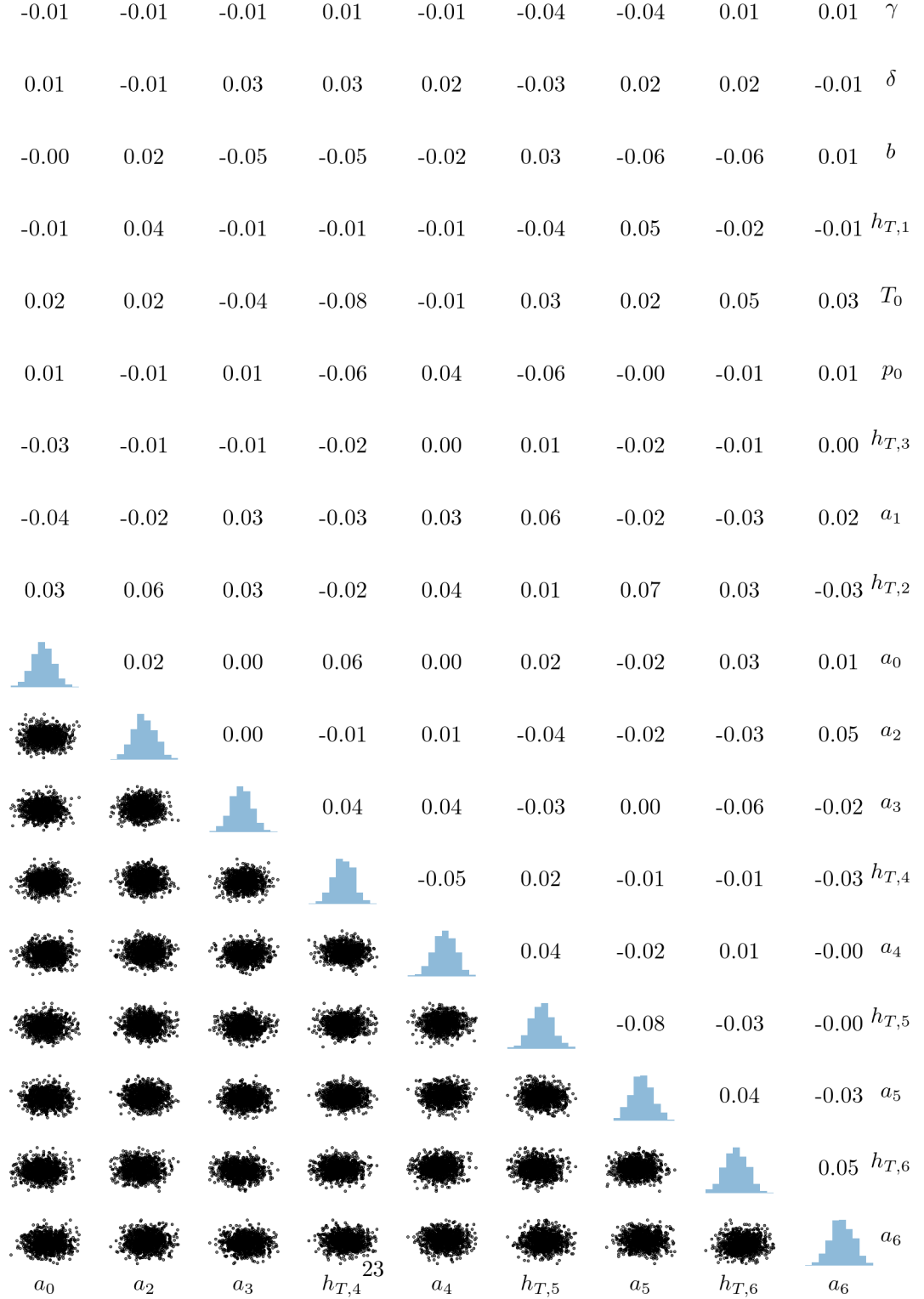


Figure 7: Plot of 1000 independent samples from TT approximation of  $\sqrt{\pi(\boldsymbol{\theta}, \lambda, \gamma | \mathbf{y})}$  via SIRT-MH scheme. We plot the Pearson correlation coefficient ranging from  $-1$  to  $1$  for each hyper-parameter pair.



Correlation plot of samples from TT-approximation of  $\sqrt{\pi(\boldsymbol{\theta}, \lambda, \gamma | \mathbf{y})}$  via SIRT-MH scheme.

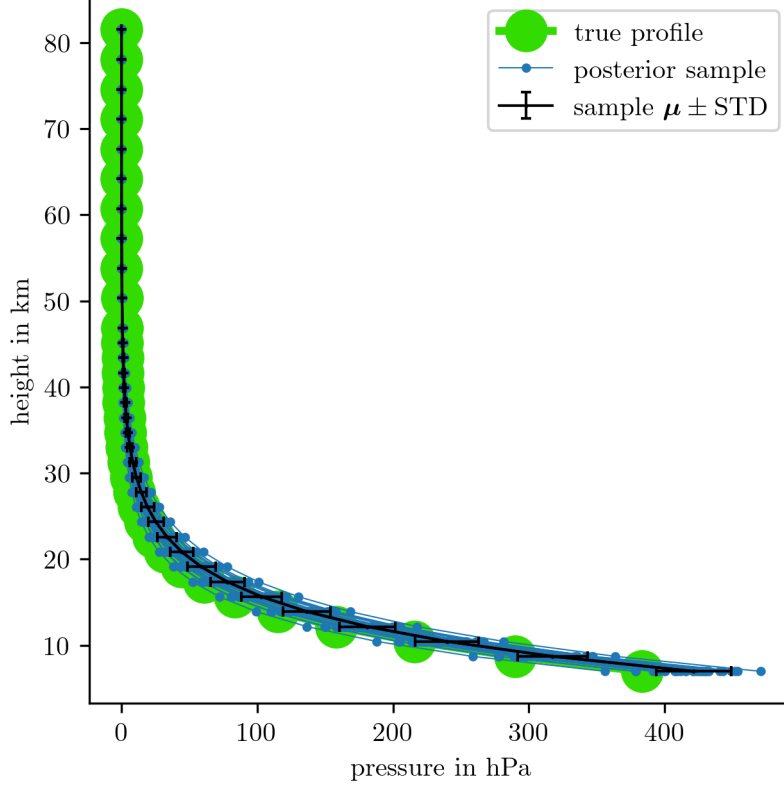


Figure 8: 29 pressure samples.

### 6.1.2 Randomise then Optimise Method to Draw Samples from the Full Conditional Posterior

To sample an posterior ozone profile from the full conditional  $\pi(\mathbf{x}|\Phi, \delta, \gamma, \mathbf{y})$  we use the RTO method, as in [1, 2, 3, 6]. Rewrite the full conditional posterior as

$$\pi(\mathbf{x}|\Phi, \delta, \gamma, \mathbf{y}) \propto \pi(\mathbf{y}|\mathbf{x}, \Phi, \gamma)\pi(\mathbf{x}|\delta) \quad (45)$$

$$\propto \exp\left(-\frac{1}{2}(\mathbf{A}_\Phi \mathbf{x} - \mathbf{y})^T \Sigma^{-1}(\mathbf{A}_\Phi \mathbf{x} - \mathbf{y})\right) \exp\left(-\frac{1}{2}(\boldsymbol{\mu} - \mathbf{x})^T \mathbf{Q}(\boldsymbol{\mu} - \mathbf{x})\right), \quad (46)$$

$$= \exp\left(-\frac{1}{2}\|\hat{\mathbf{A}}\mathbf{x} - \hat{\mathbf{y}}\|_{L^2}^2\right), \quad (47)$$

where

$$\hat{\mathbf{A}} := \begin{bmatrix} \Sigma^{-1/2} \mathbf{A}_\Phi \\ \mathbf{Q}^{1/2} \end{bmatrix}, \quad \hat{\mathbf{y}} := \begin{bmatrix} \Sigma^{-1/2} \mathbf{y} \\ \mathbf{Q}^{1/2} \boldsymbol{\mu} \end{bmatrix}, \quad (48)$$



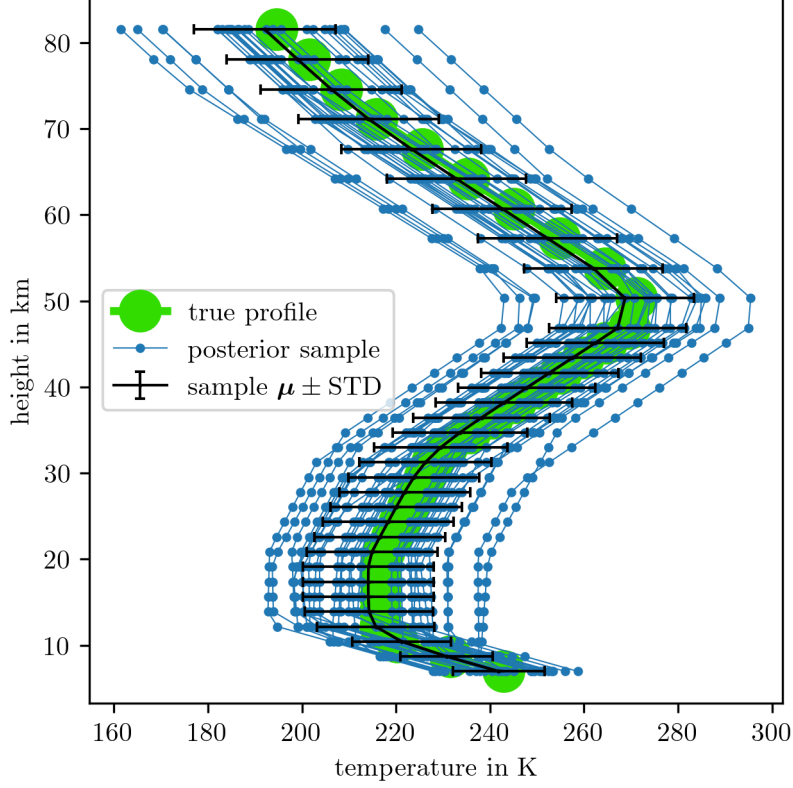


Figure 9: 29 temperature samples.

$\mathbf{Q} = \delta \mathbf{L}$  is the prior precision,  $\boldsymbol{\mu} = \mathbf{0}$  the prior mean and  $\boldsymbol{\Sigma} = \gamma^{-1} \mathbf{I}$  the noise covariance. A sample  $\mathbf{x}^{(k)}$  from the full conditional posterior  $\pi(\mathbf{x} | \boldsymbol{\Phi}, \delta, \gamma, \mathbf{y})$  is obtained by minimising the following equation:

$$\mathbf{x}^{(k)} = \arg \min_{\mathbf{x}} \|\hat{\mathbf{A}} \mathbf{x} - (\hat{\mathbf{y}} + \mathbf{b})\|_{L^2}^2 \quad (49)$$

with a random additive perturbation  $\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ . This expression becomes

$$(\mathbf{A}_{\boldsymbol{\Phi}}^T \boldsymbol{\Sigma}^{-1} \mathbf{A}_{\boldsymbol{\Phi}} + \mathbf{Q}) \mathbf{x}^{(k)} = \mathbf{A}_{\boldsymbol{\Phi}}^T \boldsymbol{\Sigma}^{-1} \mathbf{y} + \mathbf{Q} \boldsymbol{\mu} + \mathbf{v}_1 + \mathbf{v}_2, \quad (50)$$

with  $\mathbf{v}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{A}_{\boldsymbol{\Phi}}^T \boldsymbol{\Sigma}^{-1} \mathbf{A}_{\boldsymbol{\Phi}})$  and  $\mathbf{v}_2 \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}^{-1})$ , representing independent Gaussian random variables [1, 6].

More explicitly, conditioned on an independent  $\lambda^{(k)}, \gamma^{(k)} \sim \pi_{\lambda, \gamma}(\cdot | \mathbf{y})$ , one

independent full conditional posterior sample is given as

$$\mathbf{x}^{(k)} = \left( \underbrace{\gamma^{(k)} \mathbf{A}_{\Phi}^T \mathbf{A}_{\Phi} + \delta^{(k)} \mathbf{L}}_{\mathbf{B}^{(k)}} \right)^{-1} \left( \gamma^{(k)} \mathbf{A}_{\Phi}^T \mathbf{y} + \sqrt{\gamma^{(k)}} \mathbf{A}_{\Phi}^T \tilde{\mathbf{v}}_1 + \sqrt{\delta^{(k)}} \mathbf{L}^{1/2} \tilde{\mathbf{v}}_2 \right) \quad (51)$$

with  $\tilde{\mathbf{v}}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ,  $\tilde{\mathbf{v}}_2 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  and  $\mathbf{L}^{1/2}$  is the Cholesky decomposition of  $\mathbf{L}$  [1]. Note that  $\mathbf{v}_1 \in \mathbb{R}^m$  and  $\mathbf{v}_2 \in \mathbb{R}^n$ . The Cholesky factorisation of  $\mathbf{B}^{(k)}$  and  $\mathbf{L}$  is obtained via the Python function `torch numpy.linalg.cholesky` and `torch scipy.linalg.cho_solve` is used to solve for  $\mathbf{x}^{(k)}$ . We draw  $n - 1 = 29$ xxx posterior samples in  $\approx 0.003$ . The full conditional posterior samples are plotted in Fig. 15 with negative ozone values set to zero. The fact that we have to deal with negative ozone values is due to the poor prior choice in  $\pi(\mathbf{x}|\delta)$ . The posterior samples are used to find an affine map.

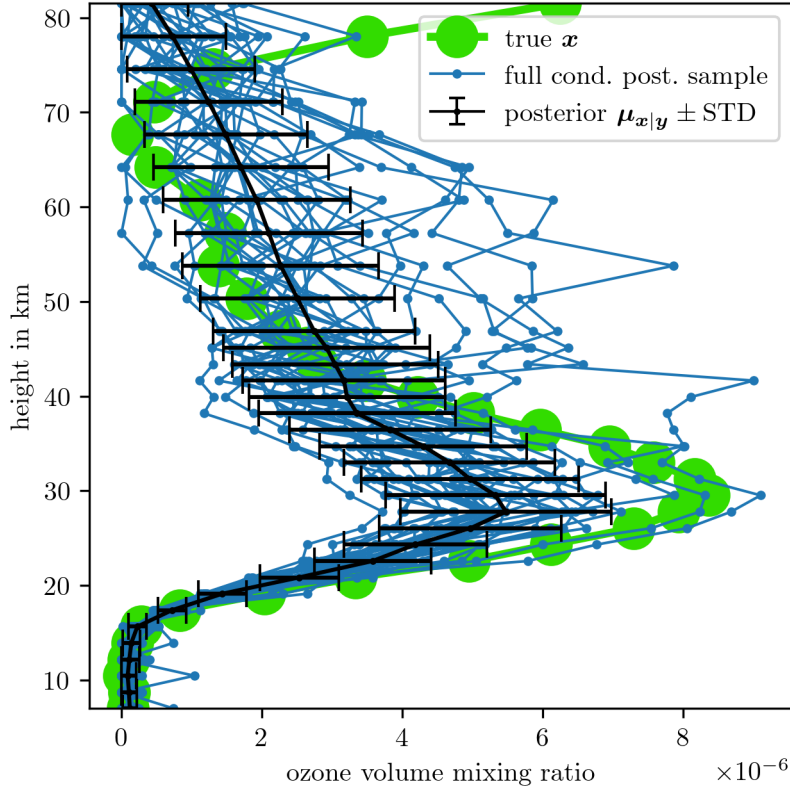


Figure 10: 29 Ozone samples via RTO method and posterior ozone mean via quadrature (see Eq. ??).

## 6.2 Finding an Affine Map

We find an affine map by creating the vector spaces  $\mathbf{W}$  based on the linear forward model and  $\mathbf{V}$  based on the non-linear forward model with ground truth pressure and temperature. More specifically  $m-1$  samples  $\mathbf{x}^{(j)} \sim \pi_{\mathbf{x}}(\cdot | \delta^{(j)}, \gamma^{(j)}, \mathbf{y})$ , for  $j = 2, \dots, m$ , from the posterior and the posterior mean  $\mu_{\mathbf{x}|\mathbf{y}}$  generate,

$$\mathbf{W} = \begin{bmatrix} \begin{array}{c} | \\ \mathbf{A}_L \mu_{\mathbf{x}|\mathbf{y}} \\ | \end{array} & \begin{array}{c} | \\ \mathbf{A}_L \mathbf{x}^{(2)} \\ | \end{array} & \dots & \begin{array}{c} | \\ \mathbf{A}_L \mathbf{x}^{(j)} \\ | \end{array} & \dots & \begin{array}{c} | \\ \mathbf{A}_L \mathbf{x}^{(m)} \\ | \end{array} \end{bmatrix} \in \mathbb{R}^{m \times m}$$

and

$$\mathbf{V} = \begin{bmatrix} \begin{array}{c} | \\ \mathbf{A}(\mu_{\mathbf{x}|\mathbf{y}}) \\ | \end{array} & \begin{array}{c} | \\ \mathbf{A}(\mathbf{x}^{(2)}) \\ | \end{array} & \dots & \begin{array}{c} | \\ \mathbf{A}(\mathbf{x}^{(j)}) \\ | \end{array} & \dots & \begin{array}{c} | \\ \mathbf{A}(\mathbf{x}^{(m)}) \\ | \end{array} \end{bmatrix} = \begin{bmatrix} - & v_1 & - \\ & \vdots & \\ - & v_j & - \\ & \vdots & \\ - & v_m & - \end{bmatrix} \in \mathbb{R}^{m \times m}.$$

Then the non-linear forward model is approximated as

$$\mathbf{A}(\mathbf{x}) \approx \mathbf{M} \mathbf{A}_L \mathbf{x}, \quad (52)$$

where we solve  $v_j = r_j \mathbf{W}$  for each row  $r_j$  in

$$\mathbf{V} \mathbf{W}^{-1} = \mathbf{M} = \begin{bmatrix} - & r_1 & - \\ & \vdots & \\ - & r_j & - \\ & \vdots & \\ - & r_m & - \end{bmatrix} \in \mathbb{R}^{m \times m}.$$

using the Python function `numpy.linalg.solve`. This is feasible since every noise-free measurement is independent of each other, and then every row  $v_j$  of  $\mathbf{V} \in \mathbb{R}^{m \times m}$  is independent of each other as well. For an  $\mathbf{x} = \mu_{\mathbf{x}|\mathbf{y}} + \Delta \mathbf{x}$  we rewrite Eq. 52 to

$$\mathbf{A}(\mathbf{x}) \approx \underbrace{\mathbf{M} \mathbf{A}_L \mu_{\mathbf{x}|\mathbf{y}}}_{=\mathbf{A}(\mu_{\mathbf{x}|\mathbf{y}})} + \underbrace{\mathbf{M} \mathbf{A}_L \Delta \mathbf{x}}_{=\mathbf{A}'(\mu_{\mathbf{x}|\mathbf{y}}) \Delta \mathbf{x}} \quad (53)$$

$$= \underbrace{\mathbf{A}'(\mu_{\mathbf{x}|\mathbf{y}}) \mathbf{x}}_{\mathbf{Ax}} + \underbrace{\mathbf{A}(\mu_{\mathbf{x}|\mathbf{y}}) - \mathbf{A}'(\mu_{\mathbf{x}|\mathbf{y}}) \mu_{\mathbf{x}|\mathbf{y}}}_{\mathbf{b}} \quad (54)$$

to show that  $\mathbf{M} : \mathbf{A}_L \mathbf{x} \rightarrow \mathbf{A}(\mathbf{x})$  is an affine map.

The relative RMS difference  $\|\text{vec}(\mathbf{M} \mathbf{W}) - \text{vec}(\mathbf{V})\|_{L^2} / \|\text{vec}(\mathbf{M} \mathbf{W})\|_{L^2}$  between the mapped linear noise-free data and the non-linear noise-free data is approximately 0.001%. This is much smaller than the relative RMS difference

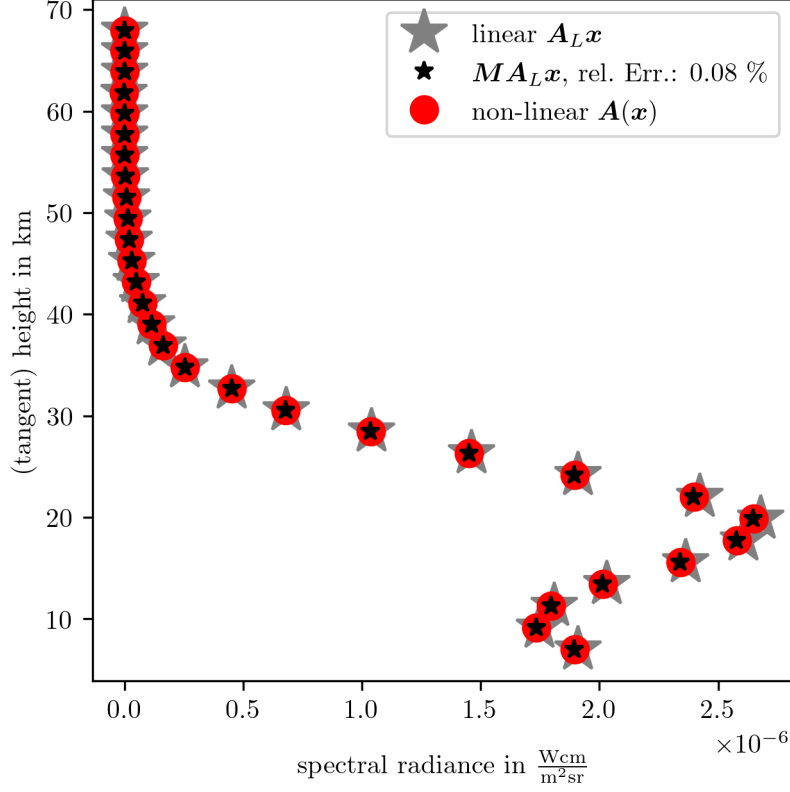
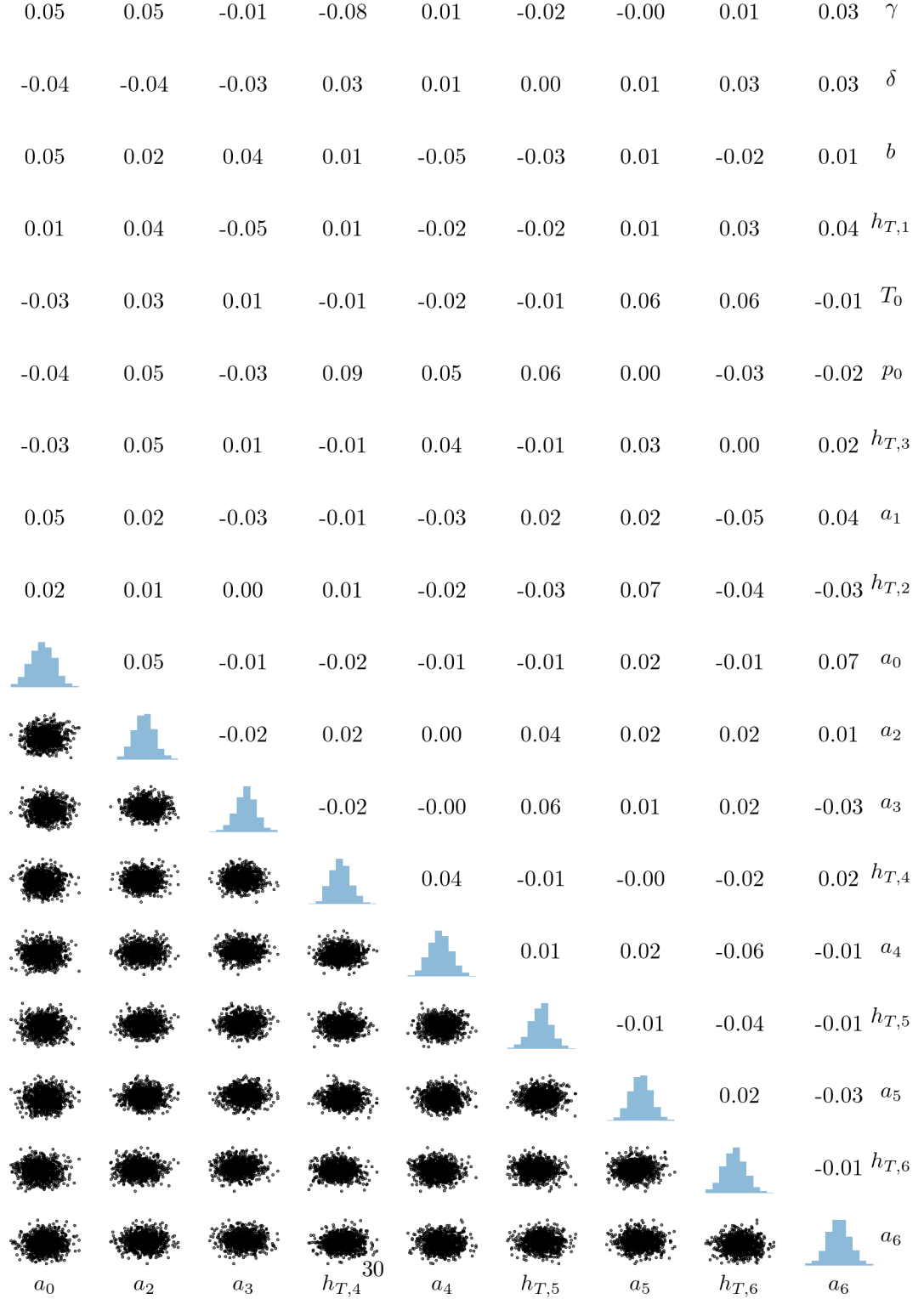


Figure 11: Assessment of how well the affine map  $\mathbf{M}$  approximates noise-free non-linear data  $\mathbf{A}(\mathbf{x})$  (red circles) from noise-free linear data  $\mathbf{A}_L \mathbf{x}$  (grey stars). The approximated noise-free data (black stars) has a relative RMS error of  $\approx 0.01\%$  compared to the true non-linear noise-free data. The ozone profile  $\mathbf{x}$  to generate this noise-free data has not been used to create the affine map.

between  $\mathbf{W}$  and  $\mathbf{V}$  of about 1%. Here  $\text{vec}(\mathbf{V})$  vectorises the matrix  $\mathbf{V}$ . Fig. 11 shows the mapping for one posterior ozone sample with a relative RMS error  $\approx 0.01\%$ . This posterior ozone sample has not been used to create this mapping; in other words, this is an unseen event not occurring in the training data. Consequently, from here onwards the approximated forward map is used. This takes  $\approx 0.1\text{s}$ .

### 6.3 Approximated Forward Model





Correlation plot of samples from TT-approximation of  $\sqrt{\pi(\boldsymbol{\theta}, \lambda, \gamma | \mathbf{y})}$  via SIRT-MH scheme.

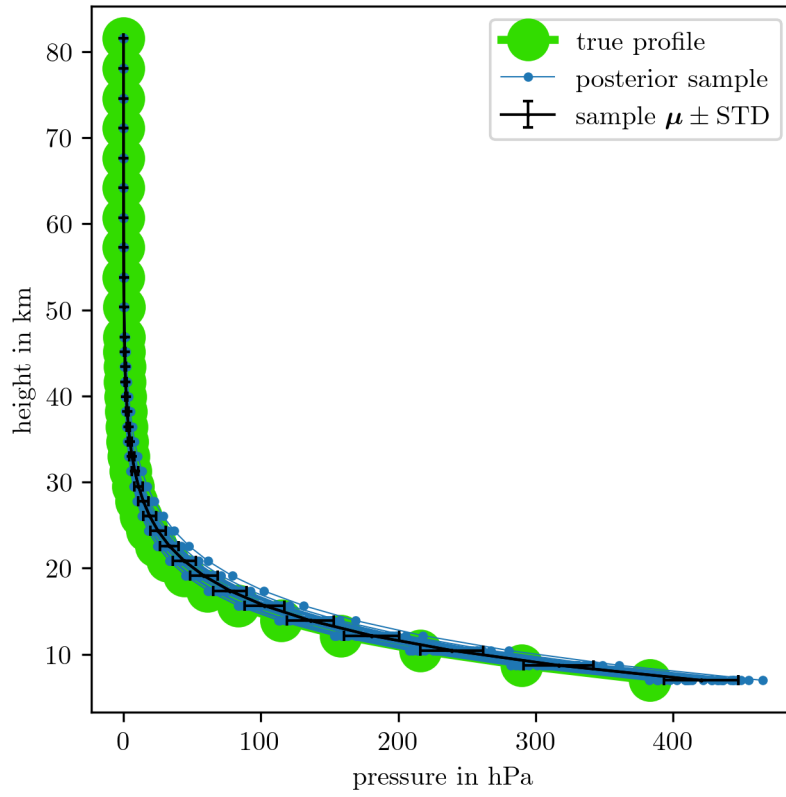


Figure 13: 29 pressure samples, approximated model.

- marginal posterior same TT setup
- conditional posterior
- plot results

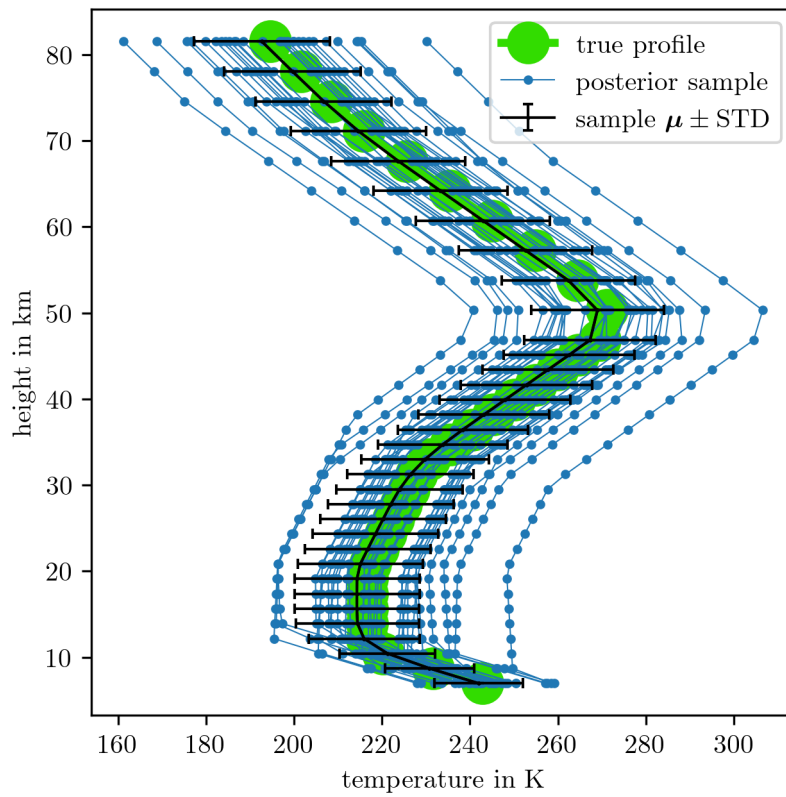


Figure 14: 29 temperature samples, approximated model.



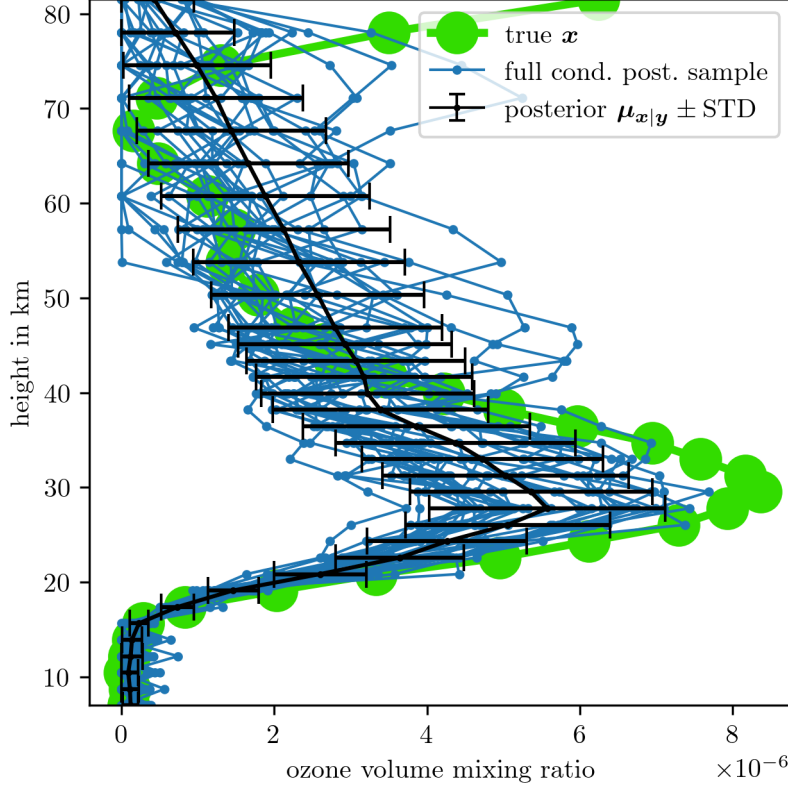


Figure 15: 29 Ozone samples via RTO method and posterior ozone mean via quadrature (see Eq. ??) with approximated model.

## 7 Conclusion

- TT approximation vs conventional sampler cite Rosenthal for IACT
- Pressure Ozone correlation
- 2nd Ozone peak not included

## References

- [1] Bardsley, J. “MCMC-based image reconstruction with uncertainty quantification”. In: *SIAM Journal on Scientific Computing* 34.3 (2012), A1316–A1332.
- [2] Bardsley, J., Solonen, A., Haario, H., and Laine, M. “Randomize-then-optimize: A method for sampling from posterior distributions in nonlinear inverse problems”. In: *SIAM Journal on Scientific Computing* 36.4 (2014), A1895–A1910.
- [3] Bardsley, J. and Cui, T. “A Metropolis-Hastings-Within-Gibbs sampler for nonlinear hierarchical-Bayesian inverse problems”. In: *2017 MATRIX Annals*. Vol. 2. MATRIX Book Series. Switzerland: Springer, 2019, pp. 2–12.
- [4] Carlotti, M. and Ridolfi, M. “Derivation of temperature and pressure from submillimetric limb observations”. In: *Applied Optics* 38.12 (Apr. 1999), pp. 2398–2409.
- [5] Champ, C. W. and Sills, A. V. “The generalized law of total covariance”. In: *preprint* (2022).
- [6] Fox, C. and Norton, R. A. “Fast sampling in a linear-Gaussian inverse problem”. In: *SIAM/ASA Journal on Uncertainty Quantification* 4.1 (2016), pp. 1191–1218.
- [7] Froidevaux, L. et al. “Validation of Aura Microwave Limb Sounder stratospheric ozone measurements”. In: *Journal of Geophysical Research: Atmospheres* 113.D15S20 (2008).
- [8] Gordon, I. E et al. “The HITRAN2020 molecular spectroscopic database”. In: *Journal of Quantitative Spectroscopy and Radiative Transfer* 277.107949 (2022).
- [9] Hesse, D. *py-uwerr; Python implementation of Monte Carlo error analysis a la Wolff*. <https://github.com/dhesse/py-uwerr>. [Online; accessed 09/09/25].
- [10] Kaipio, J. P. and Somersalo, E. *Statistical and Computational Inverse Problems*. New York: Springer-Verlag New York, 2005.
- [11] Lee, J. N. and Wu, D. L. “Solar cycle modulation of nighttime ozone near the mesopause as observed by MLS”. In: *Earth and Space Science* 7.4 (2020).
- [12] Livesey, N. J. et al. “Validation of Aura Microwave Limb Sounder O<sub>3</sub> and CO observations in the upper troposphere and lower stratosphere”. In: *Journal of Geophysical Research: Atmospheres* 113.D15S02 (2008).
- [13] Read, W., Shippony, Z., Schwartz, M., Livesey, N. J., and Van Snyder, W. “The clear-sky unpolarized forward model for the EOS aura microwave limb sounder (MLS)”. In: *IEEE Transactions on Geoscience and Remote Sensing* 44.5 (2006), pp. 1367–1379.

- [14] Readings, C. *Envisat MIPAS An Instrument for Atmospheric Chemistry and Climate Research*. Noordwijk: ESA Publications Division, 2000.
- [15] Ridolfi, M. et al. “Optimized forward model and retrieval scheme for MIPAS near-real-time dataprocessing”. In: *Applied Optics* 39.8 (Mar. 2000), pp. 1323–1340.
- [16] Rue, H. and Held, L. *Gaussian Markov Random Fields: Theory and Applications*. London: CRC Press, 2005.
- [17] Rybicki, G. B. and Lightman, A. P. *Radiative Processes in Astrophysics*. Weinheim: Wiley-VCH, 2004.
- [18] Schwartz, M., Froidevaux, L., Livesey, N., and Read, W. *MLS/Aura Level 2 Ozone (O3) Mixing Ratio V005*. [https://disc.gsfc.nasa.gov/datasets/ML203\\_005/summary?keywords=mls\%20o3](https://disc.gsfc.nasa.gov/datasets/ML203_005/summary?keywords=mls\%20o3). [Online; accessed 25/04/24]. NASA Goddard Earth Sciences Data and Information Services Center, 2020.
- [19] Šimečková, M., Jacquemart, D., Rothman, L. S., Gamache, R. R., and Goldman, A. “Einstein A-coefficients and statistical weights for molecular absorption transitions in the HITRAN database”. In: *Journal of Quantitative Spectroscopy and Radiative Transfer* 98.1 (2006), pp. 130–155.
- [20] Simpson, D., Lindgren, F., and Rue, H. “Think continuous: Markovian Gaussian models in spatial statistics”. In: *Spatial Statistics* 1 (2012), pp. 16–29.
- [21] *U.S. Standard Atmosphere, 1976*. Washington, D.C.: United States. National Oceanic and Atmospheric Administration, United States Committee on Extension to the Standard Atmosphere, 1976.
- [22] Wang, Y.-X., Sharpnack, J., Smola, A. J., and Tibshirani, R. J. “Trend filtering on graphs”. In: *Journal of Machine Learning Research* 17.105 (2016), pp. 1–41.
- [23] Waters, J. et al. “The earth observing system microwave limb sounder (EOS MLS) on the Aura satellite”. In: *IEEE Transactions on Geoscience and Remote Sensing* 44.5 (2006), pp. 1075–1092.
- [24] Wolff, U. “Monte Carlo errors with less errors”. In: *Computer Physics Communications* 156.2 (2004), pp. 143–153.
- [25] Wolff, U., Bunk, B. hard, Korzec, T., Knechtli, F., and Bär, O. *Lecture Notes on Computational Physics II [in german]*. <https://www.physik.hu-berlin.de/de/com/teachingandseminars/previousCPII>. [Online; accessed 29/08/25]. Humboldt University, Berlin, 2016.