Tensor-Train Paper

Lennart

November 2025

Abstract

The radiative transfer equation (RTE) for a stratospheric limb-sounder measuring thermal radiation of ozone presents a weakly non-linear forward map. Given some simulated data, this paper presents a hierarchical Bayesian framework to infer pressure, temperature and ozone concentration values and some other hyper-parameters such as the noise covariance and the ozone smoothness. Thereby the non-linear forward model is approximated with an affine map and the linearised RTE, which in combination with the marginal-then-conditional scheme [12] and a tensor-train approximation of the marginal posterior over the hyper-parameters enables efficient inference.

1 Introduction

There are satellites, which orbit around the earth at a height of around 500km above the ground, that carry measurement devices to determine trace gas concentrations in the stratosphere. More specifically, atmospheric limb-sounders are pointing through the atmosphere and detect thermal radiation of trace gases, e.g., ozone. Examples of such limb-sounders are the Microwave Limb Sounder (MLS) on NASA's Aura mission [35] and the Michelson Interferometer for Passive Atmospheric Sounding (MIPAS) on ESA's Envisat [25].

Measurements of such devices can be described by the radiative transfer equation (RTE), which in this case is a path integral along the satellite's pointing direction. This path integral includes an absorption term accounting for the reattenuation of the thermal radiation along the satellite's line of sight. That makes inferring the trace gas concentration from a set of measurements a non-linear inverse problem.

Existing frameworks use regularisation methods to retrieve trace gas concentration from measurements [27, 19, 22]. These methods do not include hyper-parameters, such as noise covariance, and produce biased results [13]. Additionally, atmospheric states such as pressure and temperature are retrieved from some previous measurements. Conditioned on those estimates further analysis is carried out to, e.g., retrieve ozone concentrations [19, 18].

In this paper, we develop a framework where pressure and temperature are treated as unknowns and are included within the modelling and inversion pro-

cess. In doing so, the pressure and temperature profile is parametrised, whereas a non-parametric model for the ozone is used. Firstly, given some simulated data, we treat this non-linear problem as a linear problem by neglecting the absorption term in the RTE. A linear-Gaussian hierarchical Bayesian framework is employed to infer the ozone concentration in the stratosphere and to provide pressure and temperature values. This includes establishing a hierarchical structure and classifying hyper-parameters and parameters. For efficient inference we employ the marginal-then-conditional (MTC) scheme as in [12] and extend it to include model describing hyper-parameters related to pressure and temperature. This gives a marginal posterior probability distribution over the hyper-parameters and a high-dimensional conditional posterior probability for the ozone parameter. We show that approximating the marginal posterior on a grid using a functional tensor-train (TT) is an efficient alternative to conventional sampling-based methods to generate hyper-parameter samples from the marginal posterior. Conditioned on hyper-parameter samples, ozone samples from the full conditional posterior are drawn via the randomise-then-optimise (RTO) scheme [2]. Using the results of the linearised problem, an affine map approximating the non-linear forward model is obtained. We employ the same hierarchical Bayesian framework as previously used, but with the approximated forward model, to quantify the posterior mean and variance of pressure, temperature and ozone.

All programming and analysis in this paper 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. The code will be/is available here: `https://deeptransport.github.io/deep-tensor-py/examples/`.

2 Hierarchical Bayesian Modelling

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

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

based on the forward model $\boldsymbol{A}(\boldsymbol{x})$ with an unknown parameter vector $\boldsymbol{x}$ 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 $\boldsymbol{x}$ given some data $\boldsymbol{y}$. Further, a hierarchical Bayesian model incorporates (auxiliary) hyper-parameters $\boldsymbol{\theta}$ (see Fig. 1 for a schematic representation). Within a Bayesian approach all unknown hyper-parameters and parameters are treated as random variables [16, Chapter 3].

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

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

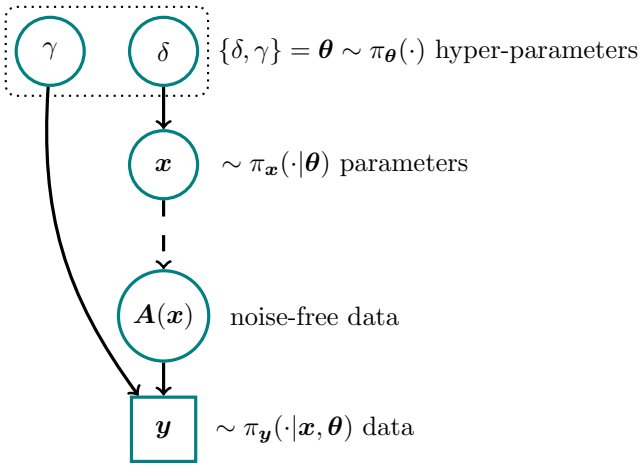


Figure 1: A directed acyclic graph (DAG) for an inverse problem visualises statistical dependencies as solid line arrows and deterministic dependencies as dotted arrows. The hyper-parameters $\boldsymbol{\theta}$ are distributed as ($\sim$) the hyper-prior distribution $\pi(\boldsymbol{\theta})$. The prior distribution $\pi_{\boldsymbol{x}}(\cdot|\boldsymbol{\theta})$ for the parameter $\boldsymbol{x}$ and the noise $\boldsymbol{\eta} \sim \pi_{\boldsymbol{\eta}}(\cdot|\boldsymbol{\theta})$ are statistically dependent on some of those hyper-parameters. Then a parameter $\boldsymbol{x} \sim \pi_{\boldsymbol{x}}(\cdot|\boldsymbol{\theta})$ is deterministically mapped through the forward model $\boldsymbol{A}(\boldsymbol{x})$. Based on the noise-free data we observe (square box) a data set $\boldsymbol{y} = \boldsymbol{A}(\boldsymbol{x}) + \boldsymbol{\eta}$ with some additive random noise, which determines the likelihood function $\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$.

with finite and non-zero $\pi(\boldsymbol{y})$. The likelihood function $\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$ is defined by the nature of the noise and the noise-free data $\boldsymbol{A}(\boldsymbol{x})$, which we read as the distribution over $\boldsymbol{y}$ conditioned on $\boldsymbol{x}$ and $\boldsymbol{\theta}$. Here $\boldsymbol{\theta}$ describe multiple hyper-parameters, e.g. the noise precision so that $\boldsymbol{\eta} \sim \pi_{\boldsymbol{\eta}}(\cdot|\boldsymbol{\theta})$, where $\sim$ reads as "is distributed as". Further, $\boldsymbol{\theta}$ may account for some physical properties of $\boldsymbol{x}$ such as the smoothness (see Sec. 4). Because all unknown parameter are treated as random variables the joint prior distribution is introduced as $\pi(\boldsymbol{x}, \boldsymbol{\theta}) = \pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ with the parameter prior distribution $\pi(\boldsymbol{x}|\boldsymbol{\theta})$ and the hyper-prior distribution $\pi(\boldsymbol{\theta})$. 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:

$$\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta} \sim \pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) \tag{3a}$$

$$\boldsymbol{x}|\boldsymbol{\theta} \sim \pi(\boldsymbol{x}|\boldsymbol{\theta}) \tag{3b}$$

$$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})\,. \tag{3c}$$

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

is defined as

$$\mathrm{E}_{\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}}[h(\boldsymbol{x})] = \int\int h(\boldsymbol{x})\,\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})\;\mathrm{d}\boldsymbol{x}\;\mathrm{d}\boldsymbol{\theta}\,. \tag{4}$$

2.1 Marginal-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 $\boldsymbol{x}$ and hyper-parameters $\boldsymbol{\theta}$, as discussed by Rue and Held in [28].

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

$$\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}) = \pi(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})\,\pi(\boldsymbol{\theta}|\boldsymbol{y}) \tag{5}$$

into the full conditional posterior $\pi(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})$ over the latent field $\boldsymbol{x}$ and the marginal posterior $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ over hyper-parameter $\boldsymbol{\theta}$. This approach, known as the marginal-and-then-conditional (MTC) method [12], is particularly advantageous when $\boldsymbol{x} \in \mathbb{R}^n$ is high-dimensional, while $\boldsymbol{\theta}$ is low-dimensional and the evaluation of the marginal posterior

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

as in [12, Lemma 2] is relatively cheap. Applying the law of total expectation [8], Eq. (4) becomes

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}}[h(\boldsymbol{x})] &= \int\int h(\boldsymbol{x})\pi(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})\;\mathrm{d}\boldsymbol{x}\,\pi(\boldsymbol{\theta}|\boldsymbol{y})\;\mathrm{d}\boldsymbol{\theta} && (7)\\
&= \int \mathbb{E}_{\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y}}\left[h(\boldsymbol{x})\right]\,\pi(\boldsymbol{\theta}|\boldsymbol{y})\;\mathrm{d}\boldsymbol{\theta} && (8)\\
&= \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y}}\left[\mathbb{E}_{\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y}}[h(\boldsymbol{x})]\right]\,. && (9)
\end{aligned}$$

In the case of a linear-Gaussian hierarchical Bayesian model, both the marginal distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ and the inner expectation $\mathbb{E}_{\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y}}\left[h(\boldsymbol{x})\right]$ are well defined (see Sec. 4 and [12]). 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 samples $\{(\boldsymbol{x},\boldsymbol{\theta})^{(1)},\ldots,(\boldsymbol{x},\boldsymbol{\theta})^{(k)},\ldots,(\boldsymbol{x},\boldsymbol{\theta})^{(N)}\} \sim \pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})$ one needs an independent sample from $\boldsymbol{\theta}^{(k)} \sim \pi(\boldsymbol{\theta}|\boldsymbol{y})$ first and then draws a sample from the full conditional posterior $\boldsymbol{x}^{(k)} \sim \pi(\boldsymbol{x}|\boldsymbol{\theta}^{(k)},\boldsymbol{y})$.

Note that for the affine case, where e.g., the forward model is given by $\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}$, the MTC method as in [12] is still applicable. For Gaussian noise and a Gaussian prior, the form of the posterior of the affine case does not change compared to the linear-Gaussian case, where the forward model may be given by $\boldsymbol{A}\boldsymbol{x}$.

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 [24] 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.

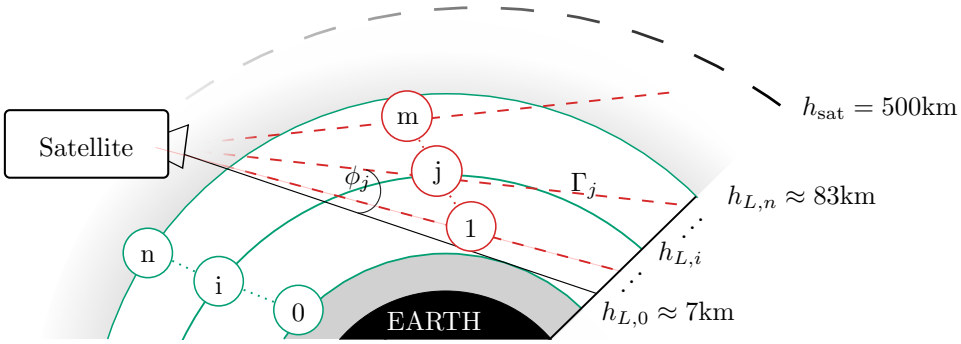


Figure 2: Schematic of measurement and analysis geometry, not to scale. The stationary satellite, at a constant height h_{sat} above Earth, takes m measurements along its line-of-sight defined by the line Γ_j. Each measurement has a pointing angle ϕ_j and a tangent height h_{ℓ_j}, $j = 1, 2, \ldots, m$ defined as the closest distance of Γ_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. 2, a satellite at a constant height h_{sat} is pointing through the atmosphere (limb-sounding) to measure thermal radiation of ozone. For each measurement $j = 1, 2, \ldots, m$, the tangent height h_{ℓ_j} and the corresponding line-of-sight Γ_j are defined. Additionally, we introduce the pointing angle $0 \leq \phi_j < \phi_{\max}$, so that if $\phi = 0\text{arc sec}$ the satellite points at $h_{L,0}$ and for a pointing angle $\phi_{\max}$ 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 $i = 1, \ldots, n$. 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) [24]

$$\int_{\Gamma_j} B(\nu,T)k(\nu,T)\frac{p(r)}{k_{\mathrm{B}}T(r)}x(r)\tau(r)\mathrm{d}r \tag{10}$$

$$\text{with } \tau(r)=\exp\Big\{-\int_{r_{\mathrm{obs}}}^{r} k(\nu,T)\frac{p(r')}{k_B T(r')}x(r')\mathrm{d}r'\Big\}. \tag{11}$$

This is a path integral along the satellite's straight line of sight Γ_j with the ozone volume mixing ratio (VMR) $x(r)$ at distance r from the satellite, at the wave number ν. Within the atmosphere, the number density $p(r)/(k_{\mathrm{B}}T(r))$ 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

$$k(\nu,T)=L(\nu,T_{\mathrm{ref}})\frac{Q(T_{\mathrm{ref}})}{Q(T)}\frac{\exp\{-c_2E''/T\}}{\exp\{-c_2E''/T_{\mathrm{ref}}\}}\frac{1-\exp\{-c_2\nu/T\}}{1-\exp\{-c_2\nu/T_{\mathrm{ref}}\}} \tag{12}$$

with Planck's constant h and speed of light c. The line intensity $L(\nu,T_{\mathrm{ref}})$ at reference temperature $T_{\mathrm{ref}}=296K$, the lower-state energy E'' in cm^{-1} of the targeted transition and the second radiation constant $c_2 \coloneqq hc/k_{\mathrm{B}} \approx 1.44\mathrm{cmK}$ are provided by the HITRAN database [14]. The total internal partition function is given as

$$Q(T)=g'\exp\{-\frac{c_2E'}{T}\}+g''\exp\{-\frac{c_2E''}{T}\}, \tag{13}$$

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 [31]), 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

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

For fundamentals on the RTE, we recommend [29, Chapter 1], and for a more comprehensive model, we refer to [23].

When simulating data, we assume an idealised limb-sounder. Since the measurement device has a negligible frequency window, the line broadening around ν for the calculations of $L(\nu,T_{\mathrm{ref}})$ is neglected. Normally, this is modelled as the convolution of the normalised Lorentz profile (collisional/pressure broadening) and the normalised Doppler profile (thermal broadening) [24]. 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 [24].

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 [30] 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. **??**, which seems to be a typical nighttime profile [17]. For more information on the processes within the atmosphere for ozone, we refer to [17].

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

$$\mathrm{d}(\log p) = \frac{\mathrm{d}p}{p} = \frac{-gM}{R^* T}\mathrm{d}h\,. \tag{15}$$

Here the acceleration due to gravity is g, the universal gas constant is $R^* = 8.31432 \times 10^{-3}\mathrm{Nm/kmol/K}$ and the mean molecular weight of the air is $M = 28.97\mathrm{kg/kmol}$, as in [33]. To enable efficient calculation of the RTE we discretise the atmosphere as in Fig. 2. Then the ozone VMR $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\} \in \mathbb{R}^n$, pressure $\boldsymbol{p} = \{p_1, p_2, \ldots, p_n\} \in \mathbb{R}^n$ and temperature $\boldsymbol{T} = \{T_1, T_2, \ldots, T_n\} \in \mathbb{R}^n$, 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 $i = 1, \ldots, n$. The hydrostatic equilibrium equation for the discretised atmosphere is

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

with $\Delta p = p_i - p_{i-1}$ and $T_{i-1} = T(h_{i-1})$ as in Eq. 18 (see also [7, 26]), for $i = 1, \ldots, n$. At sea level $h = 0\mathrm{km}$ the mean pressure is $p_0 = 1013.25\mathrm{hPa}$ and the mean temperature is $T_0 = 288.15\mathrm{K}$ [33]. The acceleration due to gravity is

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

where the polar radius of the Earth is $r_0 \approx 6356\,\mathrm{km}$, the gravitation at sea level is $g_0 \approx 9.81\mathrm{m/s}^2$. For a ground truth temperature profile we follow [33] 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

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

with gradient and height values in Tab. 1 provided by [33]. 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 [20, 35]. 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 [14]. The HITRAN database provides the line intensity $L(\nu, T_{\text{ref}})$ for the isotopologue $^{16}O_3$ with the AFGL Code 666.

To compute a data vector, we define an atmosphere between $h_{L,0} = 6.9$km and $h_{L,n} = 83.3$km with $n = 45$ equidistant layers and a satellite fixed at a height of $h_{\text{sat}} = 500$km (see Fig. 2). We measure $m = 30$ times between heights of ≈ 7km and ≈ 68km with pointing accuracy 175arc sec and equidistant spaced

pointing angles

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

Above $\approx$ 68km the data is noise dominated (see Fig. 7), hence no measurements are taken in higher altitudes. Each pointing angle ϕ_j defines a path Γ_j (see Fig. 2). The corresponding path integrals in Eq. 10 and Eq. 11 are evaluated using the trapezoidal rule and define the non-linear forward model $\boldsymbol{A}(\boldsymbol{p}, \boldsymbol{T}, \boldsymbol{x}) \in \mathbb{R}^m$ for the set of m noise-free measurements. Here, each entry A_j of $\boldsymbol{A}(\boldsymbol{p}, \boldsymbol{T}, \boldsymbol{x}) \in \mathbb{R}^m$ includes multiple evaluations of the integral in Eq. 11 to calculate the absorption $\tau(r)$. The simulated data vector

$$\boldsymbol{y} = \boldsymbol{A}(\boldsymbol{p}, \boldsymbol{T}, \boldsymbol{x}) + \boldsymbol{\eta} \tag{19}$$

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

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

where STD noise is the standard deviation of the noise. An SNR of 150 is similar to [13], 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 $\boldsymbol{A}_L(\boldsymbol{p}, \boldsymbol{T}) \in \mathbb{R}^{m \times n}$. The integral in Eq. (10) is evaluated using the trapezoidal rule and enables matrix-vector multiplication $\boldsymbol{A}_L(\boldsymbol{p}, \boldsymbol{T})\boldsymbol{x}$ to compute noise-free linear data. Since neglecting the absorption changes the measurements only slightly (about 1%, see Sec. 5.2), 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 noise-free data is given by $\boldsymbol{M}\boldsymbol{A}_L(\boldsymbol{p},\boldsymbol{T})\boldsymbol{x}$. If $\boldsymbol{M}=\boldsymbol{I}$ the forward model is described by linearised RTE as in Eq. 10. Otherwise $\boldsymbol{M}\boldsymbol{A}_L(\boldsymbol{p},\boldsymbol{T})\boldsymbol{x}$ provides an approximation to the non-linear RTE, where $\boldsymbol{M}$ is an affine map. In the following, the parameter and hyper-parameters are classified and a choice of prior distributions is established. A directed acyclic graph (DAG) is used to visualise conditional dependencies between hyper-parameters $\boldsymbol{\theta}$ and the parameter $\boldsymbol{x}$ (see Fig. 6), and how those progress through to an observation (square box) $\boldsymbol{y}$. 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 a normally distributed prior $\boldsymbol{x}|\delta \sim \mathcal{N}(0,\boldsymbol{Q}^{-1})$ 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

$$\boldsymbol{Q}=\delta\boldsymbol{L}=\delta\begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix} \tag{21}$$

which is a discrete approximation to the second derivative operator with Dirichlet boundary condition and defines a 1-dimensional Graph Laplacian as in [34, 12], accounting for smoothness in the ozone profile. We reduce the dimension of $\boldsymbol{x}$ from 45 to $n=34$ by discarding every second ozone VMR over a height of $\approx$ 47km. Doing that, while not changing $\boldsymbol{L}$ 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 $\boldsymbol{p}$ in between $h_{L,0}\approx$ 7km and $h_{L,n}\approx$ 83km can be described with an exponential function

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

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

The hyper-prior distributions for $p_0, b, T_0, \boldsymbol{h}_T, \boldsymbol{a}$ are defined to be Gaussians, and to complete the model we have to choose sensible hyper-prior variances and means. The variances of $\pi(\boldsymbol{h}_T)$ are tuned so that the temperature profile

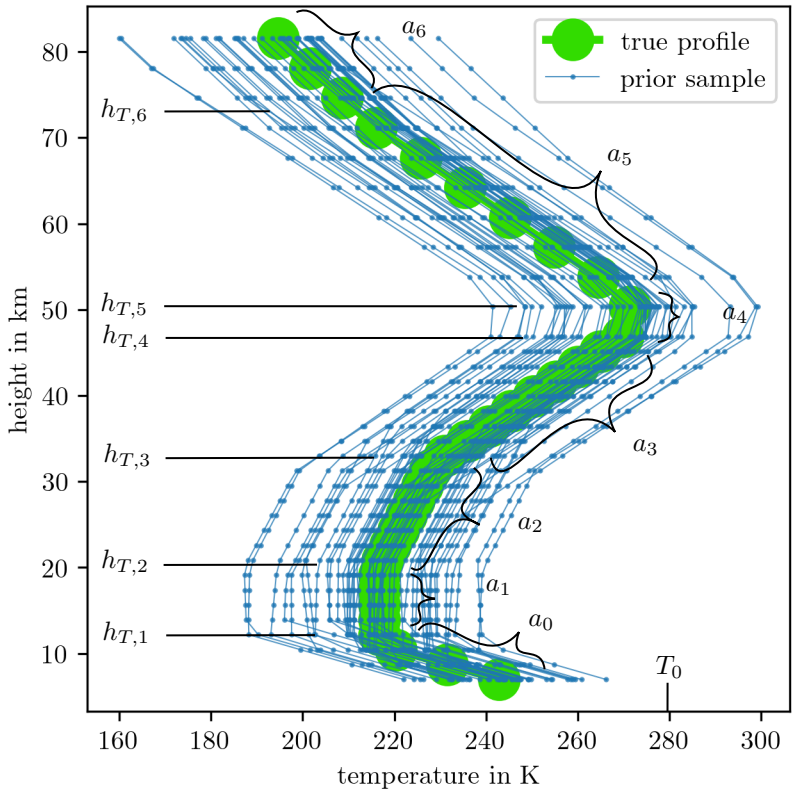


Figure 3: Prior samples from the hyper-prior distribution of $\boldsymbol{h}_T$, $\boldsymbol{a}$ and T_0, as defined in Tab. 2, where we calculate $\boldsymbol{T}$ according to the function in Eq. 18.

maintains its structure and $h_{T,i} < h_{T,i+1}$, for $i = 1, \ldots, 5$. The means of $\pi(\boldsymbol{h}_T)$ and $\pi(\boldsymbol{a})$ are set to ground truth values see Tab. 1 and the variances of $\pi(\boldsymbol{a})$ 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 [33]. These hyper-prior distributions are rather informative, because we find that the data and the model (see Fig. 5) 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 set to $\approx$ 80hPa 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 $\boldsymbol{p}$ are plotted in Fig. 4, of the temperature $\boldsymbol{T}$ in Fig. 3 and the ratio $\boldsymbol{p}/\boldsymbol{T}$ in Fig. 5. In Fig. 5 we already observe that $\boldsymbol{p}/\boldsymbol{T}$ inherits the structure of the pressure function and hence the model is uninformative about the temperature.

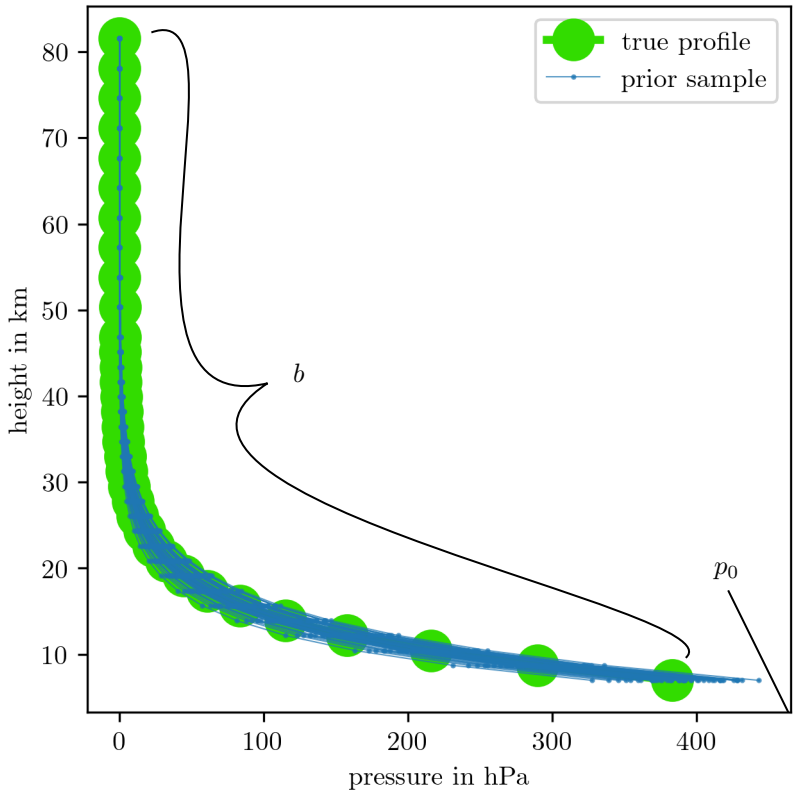


Figure 4: Prior samples from the hyper-prior distribution of b and p_0 as defined in Tab. 2, where we calculate $\boldsymbol{p}$ according to the function in Eq. 22.

For δ and γ we pick relatively uninformative Gamma distributions so that $\gamma \sim \mathcal{T}(\alpha_\gamma, \beta_\gamma) \propto \gamma^{\alpha_\gamma - 1} \exp(-\beta_\gamma \gamma)$ and $\delta \sim \mathcal{T}(\alpha_\delta, \beta_\delta)$ with $(\alpha_\gamma, \beta_\gamma) = (\alpha_\delta, \beta_\delta) = (1, 10^{-35})$ similar to [12].

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

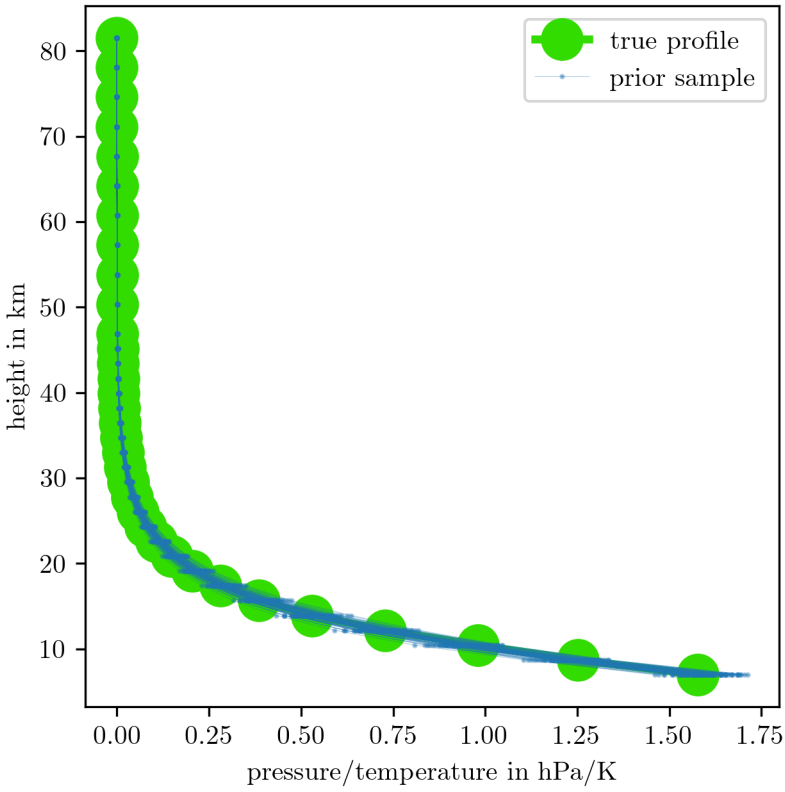


Figure 5: Prior samples from the hyper-prior distribution of $\boldsymbol{h}_T$, $\boldsymbol{a}$ and T_0 for temperature as in Eq. 18 and b and p_0 for pressure as in Eq. 22. We plot $\boldsymbol{p}/\boldsymbol{T}$. The hyper-priors are defined in Tab. 2.

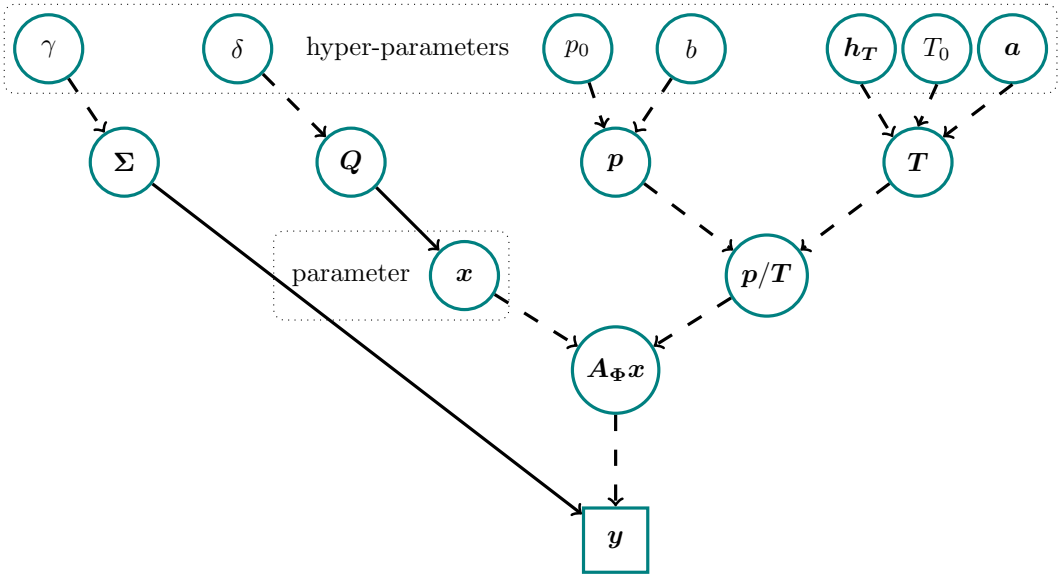


Figure 6: DAG of the hierarchical Bayesian model including ozone $\boldsymbol{x}$, pressure $\boldsymbol{p}$ and temperature $\boldsymbol{T}$. The hyper-parameters $\boldsymbol{h}_T = \{h_{T,1}, h_{T,2}, h_{T,3}, h_{T,4}, h_{T,5}, h_{T,6}\}$, $\boldsymbol{a} = \{a_0, a_1, a_2, a_3, a_4, a_5, a_6\}$, T_0, b and p_0 deterministically (dotted line) describe pressure through Eq. 22 and temperature through Eq. 18. The hyper-prior distributions $\pi(p_0, b, T_0, \boldsymbol{h_T}, \boldsymbol{a})$ are a normal distributions, see Eq. 24e to Eq. 24i. The hyper-prior distributions $\pi(\gamma)$ and $\pi(\delta)$ are gamma distributions, see Eq. 24c and Eq. 24d. The ozone parameter $\boldsymbol{x}$ is statistically (solid line) described by the prior distribution $\boldsymbol{x}|\delta \sim \mathcal{N}(0, (\delta \boldsymbol{L})^{-1})$. Here, the hyper-parameter δ accounts for smoothness in the ozone profile and defines the precision matrix $\boldsymbol{Q} = \delta \boldsymbol{L}$ as in Eq. 21. The approximated forward model $\boldsymbol{A_\Phi} \coloneqq \boldsymbol{M}\boldsymbol{A}_L(p_0, b, T_0, \boldsymbol{h_T}, \boldsymbol{a})$ with $\boldsymbol{\Phi} \coloneqq \{p_0, b, T_0, \boldsymbol{h_T}, \boldsymbol{a}\}$ describes the noise-free data $\boldsymbol{A_\Phi x}$. An observed (square box) data set $\boldsymbol{y}$ includes some additive random noise described by the noise covariance $\boldsymbol{\Sigma} = \gamma^{-1}\boldsymbol{I}$.

The DAG in Fig. 6 visualises the measurement process and conditional dependencies between the parameter $\boldsymbol{x}$ and the hyper-parameters $\boldsymbol{\theta} = \{p_0, b, T_0, \boldsymbol{h_T}, \boldsymbol{a}, \delta, \gamma\}$. This hierarchical Bayesian framework includes the hyper-parameters p_0, b for the pressure profile (see Eq. 22), $\boldsymbol{a}, \boldsymbol{h}_T, T_0$ for the temperature profile (see Eq. 18), δ for the ozone smoothness and γ for the noise precision. The hyper-parameters are described by the hyper-prior distribution $\pi(p_0, b, T_0, \boldsymbol{h_T}, \boldsymbol{a}, \delta, \gamma)$ (see Sec. 4.1). Through their respective prior distributions, pressure $\boldsymbol{p}$, temperature $\boldsymbol{T}$ and ozone $\boldsymbol{x}$ progress deterministically (dashed line) into the forward model via $\boldsymbol{x} \times \boldsymbol{p/T}$ and generate a space of all possible noise-free data Ω. Note that other variables in the RTE, such as the internal partition function and the black body radiation, are dependent on temperature as well (see Eq. 10). Finally, we observe (square box) some data $\boldsymbol{y}$ with additive normally distributed noise with zero mean and covariance $\boldsymbol{\Sigma} = \gamma^{-1}\boldsymbol{I}$. For brevity, we define the

forward model matrix as

$$\boldsymbol{A}_{\boldsymbol{\Phi}} \coloneqq \boldsymbol{M}\boldsymbol{A}_L(p_0, b, T_0, \boldsymbol{h_T}, \boldsymbol{a}) \tag{23}$$

with $\boldsymbol{\Phi} \coloneqq \{p_0, b, T_0, \boldsymbol{h_T}, \boldsymbol{a}\}$ accounting for the all pressure and temperature related hyper-parameters and $\boldsymbol{M}$ is an affine approximation. If $\boldsymbol{M} = \boldsymbol{I}$, then the forward model is based on the linearised RTE and the absorption term in the RTE, see Eq. 11, is neglected

The distributions of the hierarchical Bayesian framework are:

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

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

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

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

$$\boldsymbol{a} \sim \mathcal{N}(\boldsymbol{\mu_a}, \boldsymbol{\Sigma_a}) \tag{24e}$$

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

$$T_0 \sim \mathcal{N}(\mu_{T_0}, \sigma^2_{T_0}) \tag{24g}$$

$$p_0 \sim \mathcal{N}(\mu_{p_0}, \sigma^2_{p_0}) \tag{24h}$$

$$b \sim \mathcal{N}(\mu_b, \sigma^2_b)\,. \tag{24i}$$

Due to Gaussian noise $\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\Phi}, \delta, \gamma)$ is a normally distributed likelihood function and Eq. 24b to Eq. 24i denote prior distributions. The hyper-prior scales, shapes, means and variances are explicitly given in Tab. 2.

		TT bounds		
model parameters	priors	lower	upper	Context
$\boldsymbol{x}$	$\mathcal{N}(0, (\delta \boldsymbol{L})^{-1})$	-	-	$\boldsymbol{x}$
δ	$\mathcal{T}(1, 10^{-35})$	5×10^{10}	1×10^{13}	$\boldsymbol{x}$
γ	$\mathcal{T}(1, 10^{-35})$	8×10^{14}	6×10^{15}	$\boldsymbol{y}$
b	$\mathcal{N}(0.174, (0.01)^2)$	0.129	0.214	$\boldsymbol{p}$
$h_{T,1}$	$\mathcal{N}(11, (1.5)^2)$	5.4	16.3	$\boldsymbol{T}$
T_0	$\mathcal{N}(288.15, (10)^2)$	247	326	$\boldsymbol{T}$
p_0	$\mathcal{N}(1311, (20)^2)$	1237	1387	$\boldsymbol{p}$
$h_{T,3}$	$\mathcal{N}(32.3, (2.5)^2)$	22.9	41.7	$\boldsymbol{T}$
a_1	$\mathcal{N}(0, (0.1)^2)$	-0.38	0.38	$\boldsymbol{T}$
$h_{T,2}$	$\mathcal{N}(20.1, (0.7)^2)$	17.2	22.7	$\boldsymbol{T}$
a_0	$\mathcal{N}(-6.5, (0.01)^2)$	-6.54	-6.47	$\boldsymbol{T}$
a_2	$\mathcal{N}(1, (0.01)^2)$	0.97	1.03	$\boldsymbol{T}$
a_3	$\mathcal{N}(2.8, (0.1)^2)$	2.5	3.1	$\boldsymbol{T}$
$h_{T,4}$	$\mathcal{N}(47.4, (0.5)^2)$	45.5	49.3	$\boldsymbol{T}$
a_4	$\mathcal{N}(0, (0.1)^2)$	-0.38	0.38	$\boldsymbol{T}$
$h_{T,5}$	$\mathcal{N}(51.4, (0.5)^2)$	49.5	53.3	$\boldsymbol{T}$
a_5	$\mathcal{N}(-2.8, (0.1)^2)$	-3.18	-2.43	$\boldsymbol{T}$
$h_{T,6}$	$\mathcal{N}(71.8, (3)^2)$	60.5	83.1	$\boldsymbol{T}$
a_6	$\mathcal{N}(-2, (0.01)^2)$	-2.04	-1.96	$\boldsymbol{T}$

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.

4.2 Posterior Distribution

As explained in Sec. 2.1, we factorise the posterior

$$\pi(\boldsymbol{x}, \boldsymbol{\Phi}, \delta, \gamma|\boldsymbol{y}) \propto \pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\Phi}, \delta, \gamma)\pi(\boldsymbol{x}, \boldsymbol{\Phi}, \delta, \gamma) \tag{25}$$

into

$$\pi(\boldsymbol{x}, \boldsymbol{\Phi}, \delta, \gamma|\boldsymbol{y}) = \pi(\boldsymbol{x}|\boldsymbol{\Phi}, \delta, \gamma, \boldsymbol{y})\pi(\boldsymbol{\Phi}, \delta, \gamma|\boldsymbol{y}) \tag{26}$$

the marginal posterior $\pi(\boldsymbol{\Phi}, \delta, \gamma|\boldsymbol{y})$ and full conditional posterior $\pi(\boldsymbol{x}|\boldsymbol{\Phi}, \delta, \gamma, \boldsymbol{y})$ (see Eq. 5). As discussed in [12], for the linear-Gaussian case, $\boldsymbol{x}$ cancels in the marginal posterior over the hyper-parameters. Following the MTC scheme, we characterise the marginal posterior over the hyper-parameters $\boldsymbol{\theta} = \{\boldsymbol{\Phi}, \gamma, \delta\}$ first and then draw samples from the full conditional posterior for the parameter $\boldsymbol{x}$.

4.2.1 Marginal Posterior – Pressure and Temperature

As in [12], the marginal posterior is given as

$$\pi(\boldsymbol{\Phi}, \delta, \gamma|\boldsymbol{y}) \propto \delta^{n/2}\gamma^{m/2} \exp\left\{-\frac{1}{2}g(\boldsymbol{\Phi}, \delta, \gamma) - \frac{\gamma}{2}f(\boldsymbol{\Phi}, \delta, \gamma)\right\}\pi(\boldsymbol{\Phi}, \delta, \gamma)\,, \tag{27}$$

with,

$$f(\boldsymbol{\Phi}, \delta, \gamma) = \boldsymbol{y}^T\boldsymbol{y} - \left(\boldsymbol{A}_{\boldsymbol{\Phi}}^T\boldsymbol{y}\right)^T\left(\gamma\boldsymbol{A}_{\boldsymbol{\Phi}}^T\boldsymbol{A}_{\boldsymbol{\Phi}} + \delta\boldsymbol{L}\right)^{-1}\left(\boldsymbol{A}_{\boldsymbol{\theta}}^T\boldsymbol{y}\right), \tag{28a}$$

$$\text{and } g(\boldsymbol{\Phi}, \delta, \gamma) = \log\det\left(\gamma\boldsymbol{A}_{\boldsymbol{\Phi}}^T\boldsymbol{A}_{\boldsymbol{\Phi}} + \delta\boldsymbol{L}\right). \tag{28b}$$

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

4.2.2 Full Conditional Posterior – Ozone

As in [32], consider the joint Gaussian distribution

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

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

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

is a normal distribution and samples can be drawn via the randomise-then-optimise (RTO) method, see Sec. 5.1.2 and [2, 4, 5, 12].

5 Results

In this section, the results are presented. Some data is simulated as described in Sec. 3.2 and plotted in Fig. 7. Given the data the inverse problem is treated as a linear inverse problem by neglecting the non-linear absorption in the RTE. With the simulated data and the previously defined priors, the marginal and full conditional posterior are well defined and the MTC scheme is applied. The marginal posterior is approximated by a tensor-train (TT) and the squared inverse Rosenblatt transform (SIRT) [10] is employed to generate samples from the marginal posterior. Conditioned on those hyper-parameter samples the randomise-then-optimise (RTO) method is used to sample from the full conditional posterior. Then an affine approximation is calculated and the non-linear forward model is approximated by the affine map and the linear forward model. With the approximated forward model the just-explained methodology (TT approximation and RTO) is repeated. Finally, we present posterior ozone, pressure and temperature profiles based on the approximated forward model.

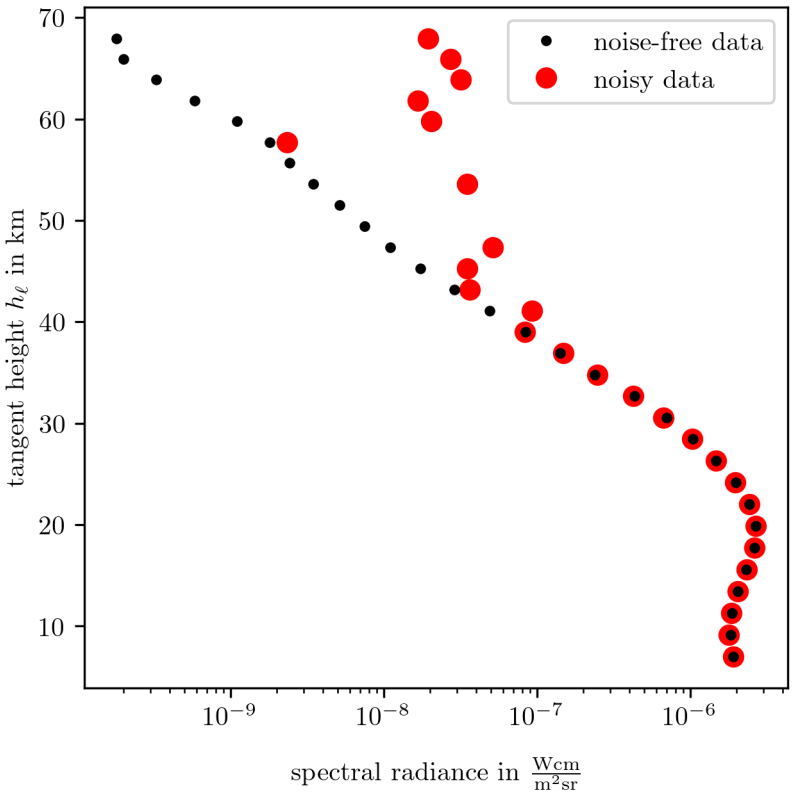


Figure 7: Non-linear noise-free data and non-linear noisy data at different tangent heights. Negative noisy data values are not shown due to logarithmic scaling.

5.1 Linear Forward Model

As already mentioned the forward model is defined as:

$$\boldsymbol{A}_{\boldsymbol{\Phi}} \coloneqq \boldsymbol{M}\boldsymbol{A}_L(\boldsymbol{\Phi}) \tag{31}$$

with $\boldsymbol{\Phi} = \{p_0, b, T_0, \boldsymbol{h_T}, \boldsymbol{a}\}$ and $\boldsymbol{M} = \boldsymbol{I}$, for now.

5.1.1 Tensor-Train Approximation of the Marginal Posterior

Instead of using conventional Markov-Chain Monte-Carlo methods, such as e.g., the t-walk [9], to generate samples from the marginal posterior, the marginal posterior is approximated directly using a TT representation. In the following a brief introduction to the TT format is provided. For more technical details we refer the reader to [11] for the inverse Rosenblatt transform (IRT), to [10] for the SIRT and to [21] for the TT-cross algorithm that generates a function approximation in the TT format.

Background on Tensor-Train Approximation Assume the target function is $\pi(\boldsymbol{\Phi}, \delta, \gamma|\boldsymbol{y})$ and the input $\boldsymbol{\theta} \coloneqq \{\boldsymbol{\Phi}, \delta, \gamma\} \in \mathbb{R}^d$ is d-dimensional. First, a d-dimensional grid with n grid points in each dimension the parameter space is defined. Usually, representing the marginal posterior naively as a d-dimensional tensor with n^d grid points is computationally not feasible. Instead a TT can be employed to approximate the target function on that parameter space with far fewer function evaluations. 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 $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ over the whole parameter space, which can be written as:

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

where each $\boldsymbol{G}_k \in \mathbb{R}^{r_{k-1} \times n \times r_k}$ 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 $k = 1, \ldots, d-1$, 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 IRT scheme [11], more precisely, the SIRT as [10] since we approximate the square root of the target function. Within this scheme a sequence of 1-dimensional functions can be constructed, so that

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

We call each function $\pi(\theta_k|\theta_{k-1},\dots,\theta_1,\boldsymbol{y})$ 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:

$$\pi(\theta_1|\boldsymbol{y}) \approx \frac{1}{z}\boldsymbol{G}_1^2(\theta_1) \int \boldsymbol{G}_2^2(\theta_2)\ \mathrm{d}\theta_2 \cdots \int \boldsymbol{G}_d^2(\theta_d)\ \mathrm{d}\theta_d, \tag{34}$$

where each of the integrals reduce a core $\boldsymbol{G}_k \in \mathbb{R}^{r_{k-1}\times n\times r_k}$ to $\int \boldsymbol{G}_k^2(\theta_k)\ \mathrm{d}\theta_k \in \mathbb{R}^{r_{k-1}\times r_k}$ and z is a normalisation constant, see [10] for more details. If you want to use a quadrature scheme to calculate expectations, e.g., $\mathbb{E}_{\pi(\boldsymbol{\theta}|\boldsymbol{y})}[f(\boldsymbol{\theta})]$, 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 θ_k the core $\boldsymbol{G}_k^2(\theta_k) \in \mathbb{R}^{r_{k-1}\times r_k}$ is given according to the used interpolation basis, see, e.g., [11], for piecewise linear interpolation (first order Lagrange polynomial).

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

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

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

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

with $u_k \in \boldsymbol{u}$. This scheme is repeated for each 'conditional marginal' in Eq. 33 until a sample $\boldsymbol{\theta} \sim \tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ is obtained.

Since the samples $\{\boldsymbol{\theta}^{(1)},\dots,\boldsymbol{\theta}^{(N)}\} \sim \tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ are drawn from an approximation they are corrected with an Metropolis–Hastings (MH) algorithm [11]. To take one MH correction step from a current state $\boldsymbol{\theta}$ to a proposed state $\boldsymbol{\theta}'$ the ratio

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

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

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

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 practice, a small constant ξ according to the L2-norm error of the TT approximation is added so that

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

This has the consequence that the chain of the MH correction algorithm is uniformly geometrically convergent (and ergodic [11]) towards $\pi(\boldsymbol{\theta}|\boldsymbol{y})$, because

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

is bounded [10]. Without ξ the TT approximation $\tilde{g}(\boldsymbol{\theta})^2$ may be arbitrarily small in regions where the true function is small and the ratio in Eq. 39 can not be bounded. With ξ we introduce a more or less uniform error. In [10] it is shown that, if ξ is chosen according to the L2-norm error

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

then the approximation satisfies the error bound

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

In theory, the SIRT scheme provides independent samples, but in practice, every system produces correlated samples. To assess this correlation, we calculate the integrated autocorrelation time (IACT) as in [36, 38] using the Python implementation from [15]. Note that we define the IACT

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

as in [12] with an autocorrelation coefficient $\Gamma(t)$. This is twice the value of the IACT in [38, pp. 103-105] and [36, 15], as commonly defined within the physics community. U. Wolff [36] (and the Python implementation by D. Hesse [15]) provide a way to not only calculate the IACT safely (choosing the summation window W) but also to quantify the errors of the estimated IACT.

Generate and use Tensor-Train Approximation To generate a TT approximation for $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ we use the `deep-tensor-py` package [6], which implements the TT-cross algorithm [21]. Further, A. de Beer [6] has implemented the SIRT scheme from [10]. By exploratory analysis we define ranks $r = [10, 10, \ldots, 2]$ manually and set the grid size $n = 30$ for the 18-dimensional parameter space, see Tab.2 for the grid bounds. Then one sweep (left-to-right and right-to-left) by the TT-cross with first order Legendre polynomials as basis functions provides an accurate enough function approximation of the marginal posterior. This requires *xxxx* function evaluations. Note that we arranged the hyper-parameter space so that highly correlated hyper-parameters are adjacent, see red boxes in Fig. 8. The Python code will be/is available here: `https://deeptransport.github.io/deep-tensor-py/examples/`. To generate a TT approximation and $\approx$ 2000 samples, see Fig. 8, from the marginal posterior including the MH-correction step takes $\approx$ 1min on a basic laptop. The maximum IACT is 1.2 ± 0.1, so every second sample via the SIRT-MH scheme is an independent sample from the marginal posterior. Conditioned on an independent marginal posterior sample $\boldsymbol{\theta}^{(k)} \sim \pi(\boldsymbol{\theta}|\boldsymbol{y})$ the RTO method can be used to generate a full conditional posterior sample $\boldsymbol{x}^{(k)} \sim \pi(\boldsymbol{x}|\boldsymbol{\theta}^{(k)}, \boldsymbol{y})$.

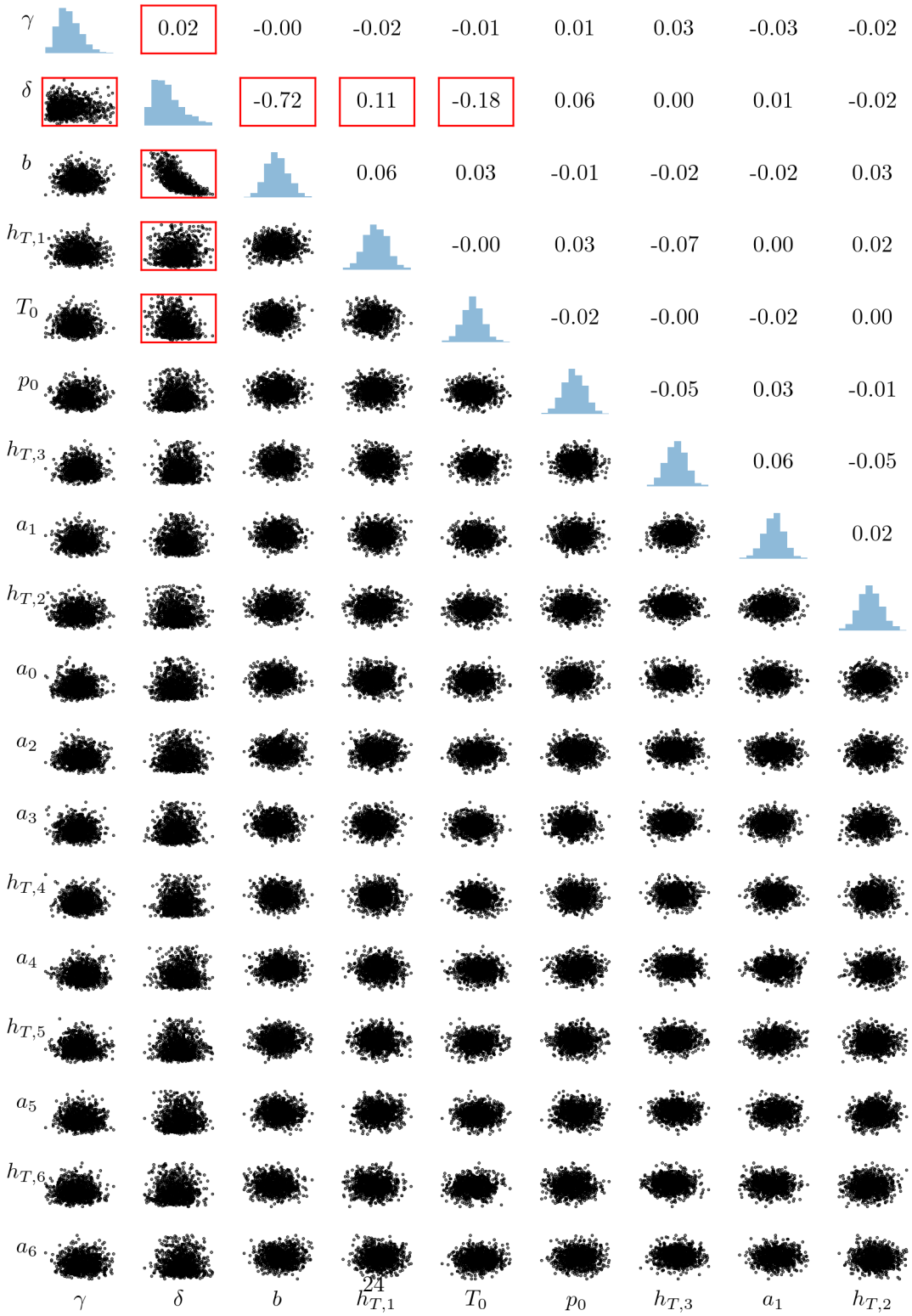
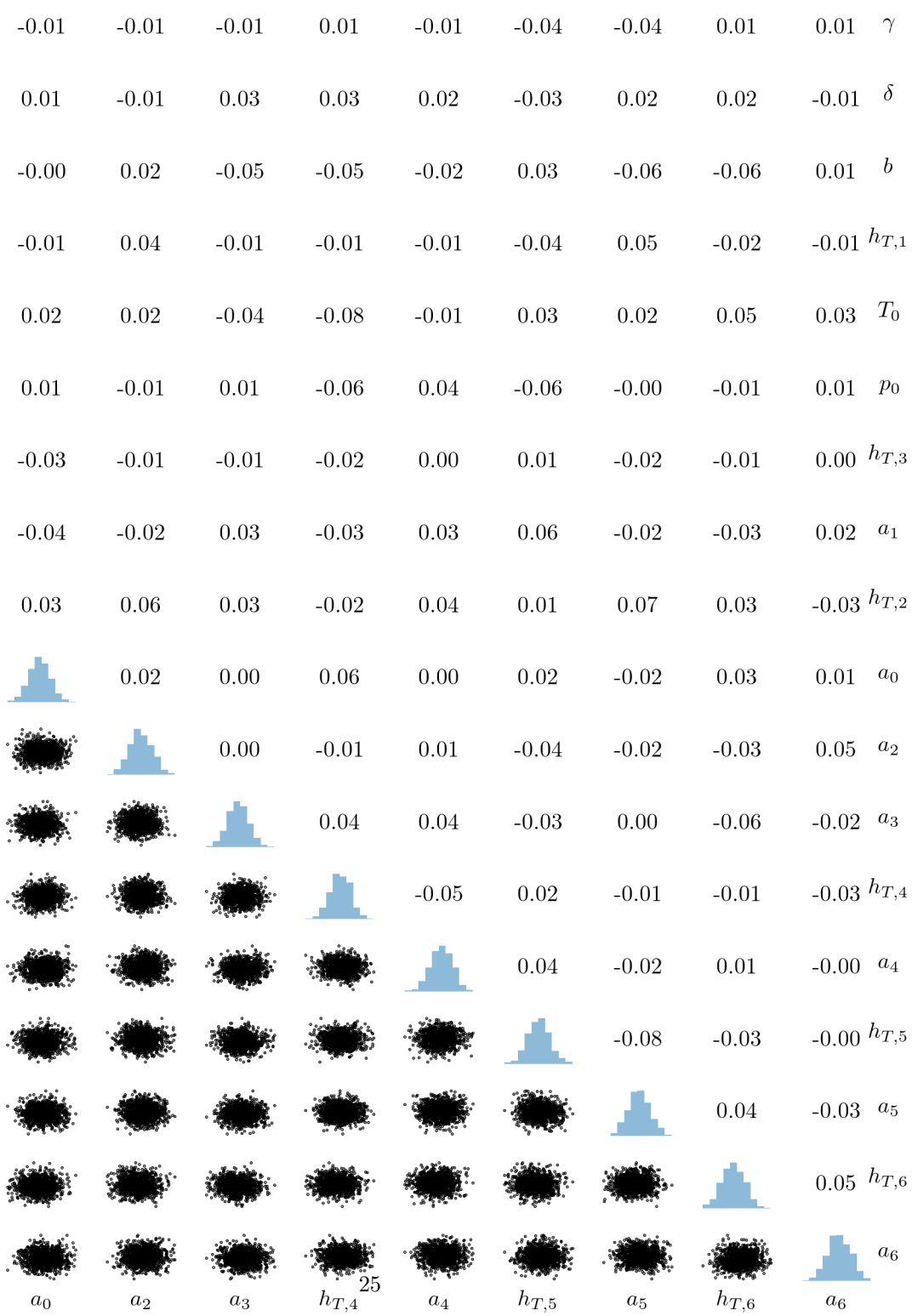


Figure 8: Plot of 1000 independent samples from the marginal posterior $\pi(\boldsymbol{\Phi}, \delta, \gamma|\boldsymbol{y})$ based on the linearised forward model via the SIRT-MH scheme. We plot the Pearson correlation coefficient ranging from -1 to 1 for each hyper-parameter pair.



Correlation plot of samples from the marginal posterior $\pi(\boldsymbol{\Phi}, \delta, \gamma|\boldsymbol{y})$ based on the linearised forward model via the SIRT-MH scheme.

5.1.2 Samples from the Full Conditional Posterior

To sample an posterior ozone profile from the full conditional $\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ with $\boldsymbol{\theta} = \{\boldsymbol{\Phi}, \delta, \gamma\}$ we use the RTO method, as in [2, 4, 5, 12]. To derive the RTO method, rewrite the full conditional posterior as

$$\pi(\boldsymbol{x}|\boldsymbol{\Phi}, \delta, \gamma, \boldsymbol{y}) \propto \pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\Phi}, \gamma)\pi(\boldsymbol{x}|\delta) \tag{43}$$

$$\propto \exp\left(-\frac{1}{2}(\boldsymbol{A}_{\boldsymbol{\Phi}}\boldsymbol{x} - \boldsymbol{y})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{A}_{\boldsymbol{\Phi}}\boldsymbol{x} - \boldsymbol{y})\right)\exp\left(-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{x})^T\boldsymbol{Q}(\boldsymbol{\mu} - \boldsymbol{x})\right), \tag{44}$$

$$= \exp\left(-\frac{1}{2}\left\|\hat{\boldsymbol{A}}\boldsymbol{x} - \hat{\boldsymbol{y}}\right\|_{L^2}^2\right), \tag{45}$$

where

$$\hat{\boldsymbol{A}} \coloneqq \begin{bmatrix}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{A}_{\boldsymbol{\Phi}} \\ \boldsymbol{Q}^{1/2}\end{bmatrix}, \quad \hat{\boldsymbol{y}} \coloneqq \begin{bmatrix}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{y} \\ \boldsymbol{Q}^{1/2}\boldsymbol{\mu}\end{bmatrix}, \tag{46}$$

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

$$\boldsymbol{x}^{(k)} = \arg\min_{\boldsymbol{x}} \|\hat{\boldsymbol{A}}\boldsymbol{x} - (\hat{\boldsymbol{y}} + \boldsymbol{b})\|_{L^2}^2 \tag{47}$$

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

$$\left(\boldsymbol{A}_{\boldsymbol{\Phi}}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A}_{\boldsymbol{\Phi}} + \boldsymbol{Q}\right)\boldsymbol{x}^{(k)} = \boldsymbol{A}_{\boldsymbol{\Phi}}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y} + \boldsymbol{Q}\boldsymbol{\mu} + \boldsymbol{v}_1 + \boldsymbol{v}_2, \tag{48}$$

with $\boldsymbol{v}_1 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{A}_{\boldsymbol{\Phi}}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A}_{\boldsymbol{\Phi}})$ and $\boldsymbol{v}_2 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}^{-1})$, representing independent Gaussian random variables [2, 12].

More explicitly, conditioned on an independent $\boldsymbol{\Phi}^{(k)}, \delta^{(k)}, \gamma^{(k)} \sim \pi(\boldsymbol{\Phi}^{(k)}, \delta^{(k)}, \gamma^{(k)}|\boldsymbol{y})$, one full conditional posterior sample is given as

$$\boldsymbol{x}^{(k)} = \Big(\underbrace{\gamma^{(k)}\boldsymbol{A}_{\boldsymbol{\Phi}}^T\boldsymbol{A}_{\boldsymbol{\Phi}} + \delta^{(k)}\boldsymbol{L}}_{\boldsymbol{B}^{(k)}}\Big)^{-1}\left(\gamma^{(k)}\boldsymbol{A}_{\boldsymbol{\Phi}}^T\boldsymbol{y} + \sqrt{\gamma^{(k)}}\boldsymbol{A}_{\boldsymbol{\Phi}}^T\tilde{\boldsymbol{v}}_1 + \sqrt{\delta^{(k)}}\boldsymbol{L}^{1/2}\tilde{\boldsymbol{v}}_2\right) \tag{49}$$

with $\tilde{\boldsymbol{v}}_1 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, $\tilde{\boldsymbol{v}}_2 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and $\boldsymbol{L}^{1/2}$ is the Cholesky decomposition of $\boldsymbol{L}$ [2]. Note that $\boldsymbol{v}_1 \in \mathbb{R}^m$ and $\boldsymbol{v}_2 \in \mathbb{R}^n$. The Cholesky factorisation of $\boldsymbol{B}^{(k)}$ and $\boldsymbol{L}$ is obtained via the Python function `torch.linalg.cholesky` and `torch.cholesky_solve` is used to solve for $\boldsymbol{x}^{(k)}$. We draw $N = 1000$ posterior samples in $\approx$ 1s. The full conditional posterior samples are plotted in Fig. 9 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 $\pi(\boldsymbol{x}|\delta)$. The posterior samples are used to find an affine map.

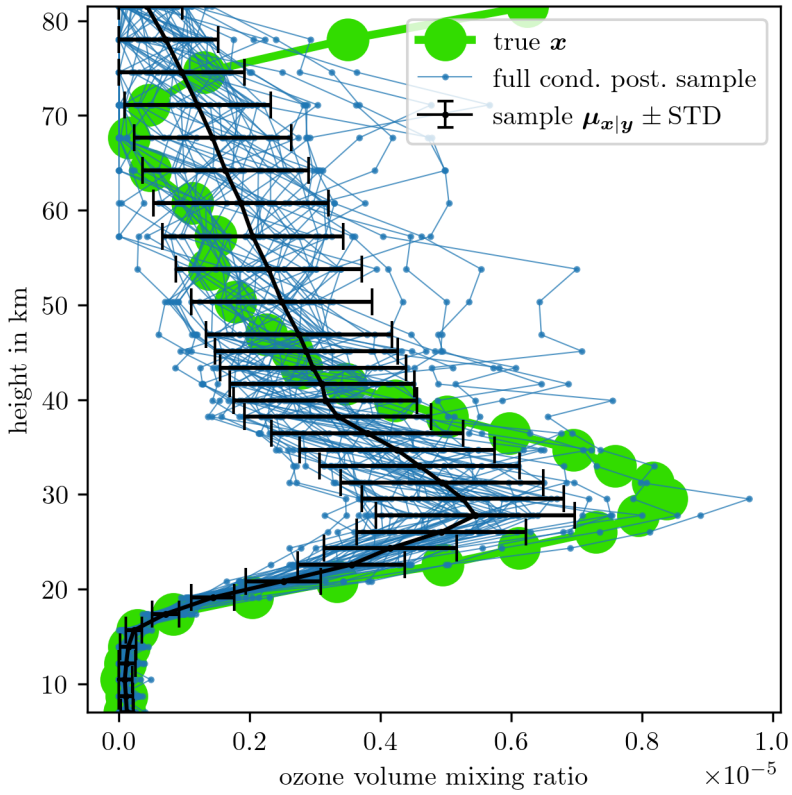


Figure 9: 50 ozone samples from the full conditional posterior via the RTO method.

5.2 Finding an Affine Map

We find an affine map by creating the vector spaces $\boldsymbol{W}$ based on the linear forward model and $\boldsymbol{V}$ based on the non-linear forward model with ground truth pressure and temperature. More specifically $N = 1000$ samples $\boldsymbol{x}^{(k)} \sim \pi(\boldsymbol{x}|\boldsymbol{\Phi}^{(k)}, \delta^{(k)}, \gamma^{(k)}, \boldsymbol{y})$, for $k = 1, \ldots, N$, generate

$$\boldsymbol{W} = \begin{bmatrix} | & & | & & | \\ \boldsymbol{A}_L(\boldsymbol{\Phi}^{(1)})\boldsymbol{x}^{(1)} & \cdots & \boldsymbol{A}_L(\boldsymbol{\Phi}^{(k)})\boldsymbol{x}^{(k)} & \cdots & \boldsymbol{A}_L(\boldsymbol{\Phi}^{(N)})\boldsymbol{x}^{(N)} \\ | & & | & & | \end{bmatrix} \in \mathbb{R}^{m \times N}$$

and

$$\boldsymbol{V} = \begin{bmatrix} | & & | & & | \\ \boldsymbol{A}(\boldsymbol{\Phi}^{(1)}, \boldsymbol{x}^{(1)}) & \cdots & \boldsymbol{A}(\boldsymbol{\Phi}^{(k)}, \boldsymbol{x}^{(k)}) & \cdots & \boldsymbol{A}(\boldsymbol{\Phi}^{(N)}, \boldsymbol{x}^{(N)}) \\ | & & | & & | \end{bmatrix} = \begin{bmatrix} — & v_1 & — \\ & \vdots & \\ — & v_j & — \\ & \vdots & \\ — & v_m & — \end{bmatrix} \in \mathbb{R}^{m \times N}.$$

Then the non-linear forward model is approximated as

$$\boldsymbol{A}(\boldsymbol{p}, \boldsymbol{T}, \boldsymbol{x}) \approx \boldsymbol{M}\boldsymbol{A}_L(\boldsymbol{p}, \boldsymbol{T})\boldsymbol{x}\,, \tag{50}$$

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

$$\boldsymbol{M} = \begin{bmatrix} — & r_1 & — \\ & \vdots & \\ — & r_j & — \\ & \vdots & \\ — & r_m & — \end{bmatrix} \in \mathbb{R}^{m \times m}.$$

using the Python function `torch.linalg.lstsq`.

The relative RMS difference $\|\mathrm{vec}(\boldsymbol{M}\boldsymbol{W}) - \mathrm{vec}(\boldsymbol{V})\|_{L^2} / \|\mathrm{vec}(\boldsymbol{V})\|_{L^2}$ between the mapped linear noise-free data and the non-linear noise-free data is approximately 0.06%. This is much smaller than the relative RMS difference between $\boldsymbol{W}$ and $\boldsymbol{V}$ of about 1%. Here $\mathrm{vec}(\boldsymbol{V})$ vectorises the matrix $\boldsymbol{V}$. Fig. 10 shows the mapping for one posterior ozone sample with a relative RMS error $\approx 0.08\%$. 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.

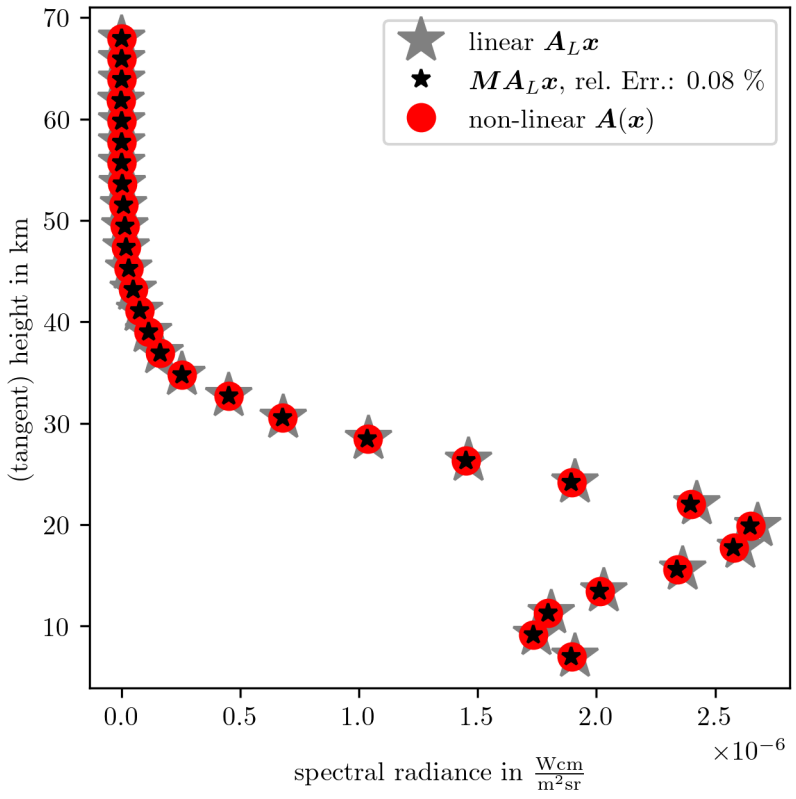


Figure 10: Assessment of how well the affine map $\boldsymbol{M}$ approximates noise-free non-linear data $\boldsymbol{A}(\boldsymbol{x})$ (red circles) from noise-free linear data $\boldsymbol{A}_L\boldsymbol{x}$ (grey stars). The approximated noise-free data (black stars) has a relative RMS error of $\approx 0.08\%$ compared to the true non-linear noise-free data. The ozone profile $\boldsymbol{x}$ to generate this noise-free data has not been used to create the affine map.

5.3 Approximated Forward Model

Now, the forward model is defined as:

$$\boldsymbol{A}_{\boldsymbol{\Phi}} \coloneqq \boldsymbol{M}\boldsymbol{A}_L(\boldsymbol{\Phi}) \tag{51}$$

with $\boldsymbol{\Phi} = \{p_0, b, T_0, \boldsymbol{h_T}, \boldsymbol{a}\}$ and $\boldsymbol{M}$ is the just obtained affine map. We repeat the exact same procedure as described in Sec. 5.1.

The max IACT is $\approx 1.2 \pm 0.1$, hence every second marginal posterior sample generated by the SIRT scheme presents an independent sample from the marginal posterior. The independent marginal posterior samples directly give posterior pressure and temperature profiles through their respective functions, see Eq. 22 and Eq. 18. The posterior pressure profiles are plotted in Fig. 11 and the posterior temperature profiles in Fig. 12. Compared to the temperature prior profiles in Fig. 3 the posterior profiles do not change. The posterior pressure profiles are slightly larger than their ground truth.

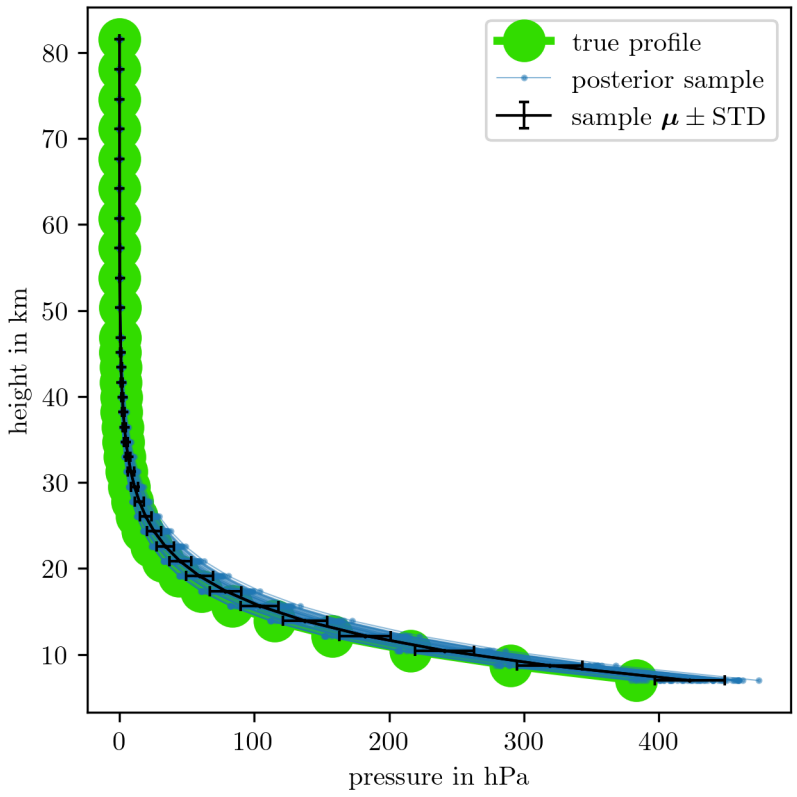


Figure 11: 50 posterior pressure profiles from marginal posterior samples based on the approximated model via Eq. 22.

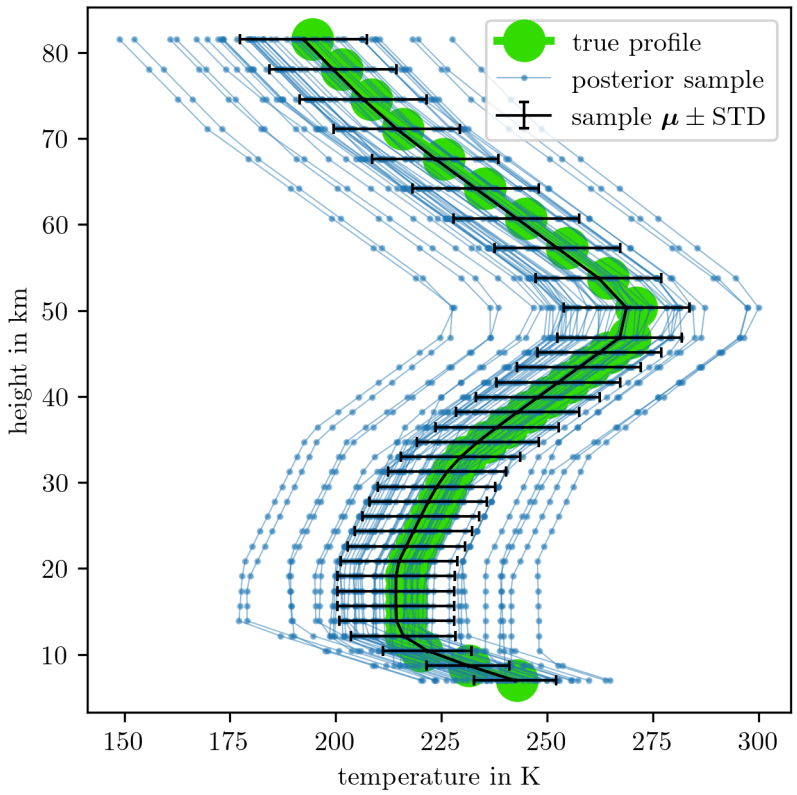


Figure 12: 50 posterior temperature profiles from marginal posterior samples based on the approximated model via Eq. 18.

Next, with the RTO method as in Sec. 5.1.2 full conditional posterior ozone samples are obtained and plotted in Fig. 13. As already indicated by Fig. 8 the pressure values and ozone concentrations are highly correlated. That is why if the pressure values are slightly larger than their ground truth, see Fig. 11, then the ozone concentration is much lower than its ground truth, see Fig. 13. Further, the posterior ozone profiles do not include a second ozone peak from the ground truth.

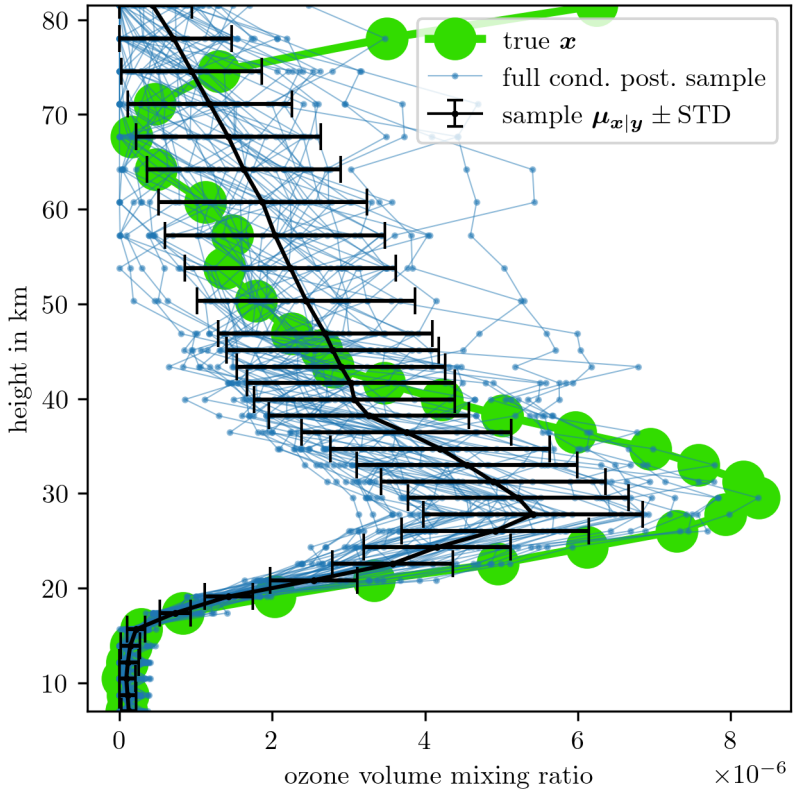


Figure 13: 50 posterior ozone samples via the RTO method based on the approximated model.

6 Conclusion

Including ozone, pressure and temperature within the inversion process, we show that ozone and pressure are highly correlated. The results show that a small change in pressure does have a big influence on the ozone profile. To allow more variability of the pressure profile, one could combine the current parametrised model of pressure with a non-parametric model, see e.g., [1]. It is clear that the data is uninformative about the ozone concentration in higher atmospheric regions and the second ozone peak is not recovered. Further we show that the model and the data do not contain information about the temperature profile.

The here employed TT framework to approximate the marginal posterior distribution provides an efficient way to generate hyper-parameter samples. To obtain an approximation of the marginal posterior xxxx function evaluations are required. Once the TT is composed every second sample generated by the SIRT scheme presents an independent sample of the marginal posterior. This is close to optimal (IACT = 1), compare to other samplers?

References

[1] Aimen, N., Maturana-Russel, P., Vajpeyi, A., Christensen, N., and Meyer, R. "Bayesian power spectral density estimation for LISA noise based on penalized splines with a parametric boost". In: *Phys. Rev. D* 113 (2 2026), p. 024022.

[2] Bardsley, J. "MCMC-based image reconstruction with uncertainty quantification". In: *SIAM Journal on Scientific Computing* 34.3 (2012), A1316–A1332.

[3] Bardsley, J., Seppanen, A., Solonen, A., Haario, H., and Kaipio, J. "Randomize-then-optimize for sampling and uncertainty quantification in electrical impedance tomography". In: *SIAM/ASA Journal on Uncertainty Quantification* 3.1 (2015), pp. 1136–1158.

[4] 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.

[5] 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.

[6] de Beer, A. *deep-tensor-py – A PyTorch implementation of the deep inverse Rosenblatt transport (DIRT) algorithm*. `https://deeptransport.github.io/deep-tensor-py/`. [Online; accessed 26/01/26]. The University of Sydney.

[7] Carlotti, M. and Ridolfi, M. "Derivation of temperature and pressure from submillimetric limb observations". In: *Applied Optics* 38.12 (Apr. 1999), pp. 2398–2409.

[8] Champ, C. W. and Sills, A. V. "The generalized law of total covariance". In: *preprint* (2022).

[9] Christen, J. A. and Fox, C. *The t-walk software*. `https://www.cimat.mx/~jac/twalk/`. [Online; accessed 25/11/24]. CIMAT, Mexico, and University of Otago, New Zealand.

[10] Cui, T. and Dolgov, S. "Deep composition of tensor-trains using squared inverse rosenblatt transports". In: *Foundations of Computational Mathematics* 22.6 (2022), pp. 1863–1922.

[11] Dolgov, S., Anaya-Izquierdo, K., Fox, C., and Scheichl, R. "Approximation and sampling of multivariate probability distributions in the tensor train decomposition". In: *Statistics and Computing* 30 (2020), pp. 603–625.

[12] 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.

[13] Froidevaux, L. et al. "Validation of Aura Microwave Limb Sounder stratospheric ozone measurements". In: *Journal of Geophysical Research: Atmospheres* 113.D15S20 (2008).

[14] Gordon, I. E et al. "The HITRAN2020 molecular spectroscopic database". In: *Journal of Quantitative Spectroscopy and Radiative Transfer* 277.107949 (2022).

[15] 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].

[16] Kaipio, J. P. and Somersalo, E. *Statistical and Computational Inverse Problems*. New York: Springer-Verlag New York, 2005.

[17] 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).

[18] Livesey, N. J. and Snyder, W. V. *Earth Observing System (EOS) Microwave Limb Sounder (MLS) Retrieval Processes Algorithm – Theoretical Basis – Version 2.0.* `https://mls.jpl.nasa.gov/data/eos_algorithm_atbd.pdf`. [Online; accessed 28/01/26]. 2004.

[19] Livesey, N. J., Van Snyder, W, Read, W. G., and Wagner, P. A. "Retrieval algorithms for the EOS Microwave limb sounder (MLS)". In: *IEEE Transactions on Geoscience and Remote Sensing* 44.5 (2006), pp. 1144–1155.

[20] Livesey, N. J. et al. "Validation of Aura Microwave Limb Sounder O3 and CO observations in the upper troposphere and lower stratosphere". In: *Journal of Geophysical Research: Atmospheres* 113.D15S02 (2008).

[21] Oseledets, I. and Tyrtyshnikov, E. "TT-cross approximation for multidimensional arrays". In: *Linear Algebra and its Applications* 432.1 (2010), pp. 70–88.

[22] Raspollini, P. et al. "Level 2 processor and auxiliary data for ESA Version 8 final full mission analysis of MIPAS measurements on ENVISAT". In: *Atmospheric Measurement Techniques Discussions* 2021 (2021), pp. 1–46.

[23] 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.

[24] Readings, C. *Envisat MIPAS An Instrument for Atmospheric Chemistry and Climate Research*. Noordwijk: ESA Publications Division, 2000.

[25] Readings, C. and Harris, R. A. *Envisat MIPAS an Instrument for Atmospheric Chemistry and Climate Research.* `https://earth.esa.int/eogateway/documents/20142/37627/envisat-mipas-instrument-description.pdf`. [Online; accessed 16/07/22]. 2000.

[26] 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.

[27] Rodgers, C. D. “Retrieval of atmospheric temperature and composition from remote measurements of thermal radiation”. In: *Reviews of Geophysics* 14.4 (1976), pp. 609–624.

[28] Rue, H. and Held, L. *Gaussian Markov Random Fields: Theory and Applications*. London: CRC Press, 2005.

[29] Rybicki, G. B. and Lightman, A. P. *Radiative Processes in Astrophysics*. Weinheim: Wiley-VCH, 2004.

[30] Schwartz, M., Froidevaux, L., Livesey, N., and Read, W. *MLS/Aura Level 2 Ozone (O3) Mixing Ratio V005*. `https://disc.gsfc.nasa.gov/datasets/ML2O3_005/summary?keywords=mls%20o3`. [Online; accessed 25/04/24]. NASA Goddard Earth Sciences Data and Information Services Center, 2020.

[31] 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.

[32] Simpson, D., Lindgren, F., and Rue, H. “Think continuous: Markovian Gaussian models in spatial statistics”. In: *Spatial Statistics* 1 (2012), pp. 16–29.

[33] *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.

[34] 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.

[35] 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.

[36] Wolff, U. “Monte Carlo errors with less errors”. In: *Computer Physics Communications* 156.2 (2004), pp. 143–153.

[37] Wolff, U. *UWerr.m Version 6*. `https://www.physik.hu-berlin.de/de/com/ALPHAsoft`. [Online; accessed 5/11/23]. Humboldt-Universität to Berlin, 2004.

[38] Wolff, U., Bunk, B., 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.

7 Additional Figures

7.1 Pressure and Temperature Function

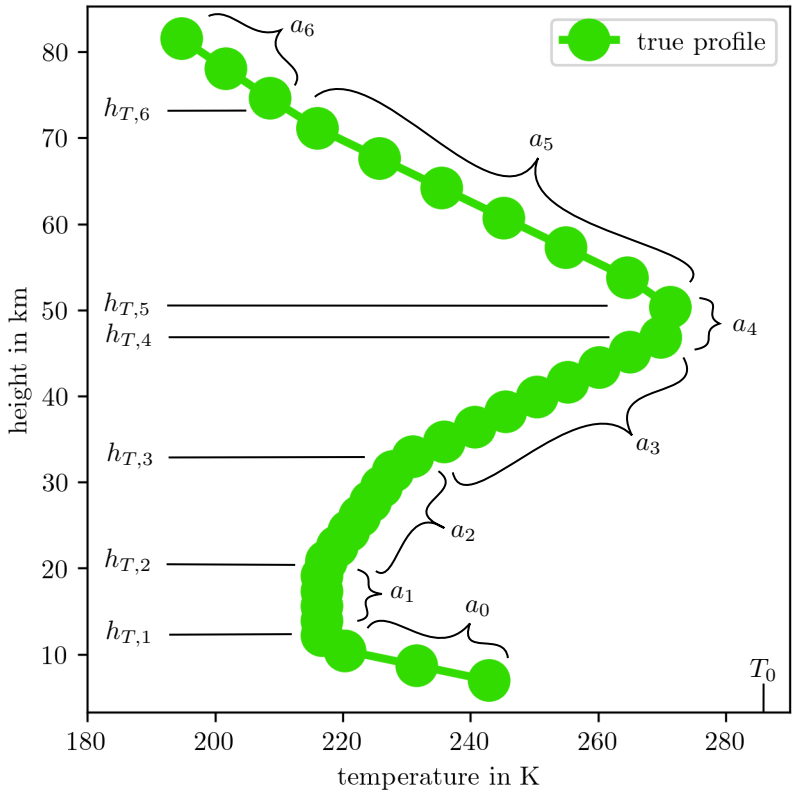


Figure 14: True temperature profile including hyper-parameters as in Eq. 18.

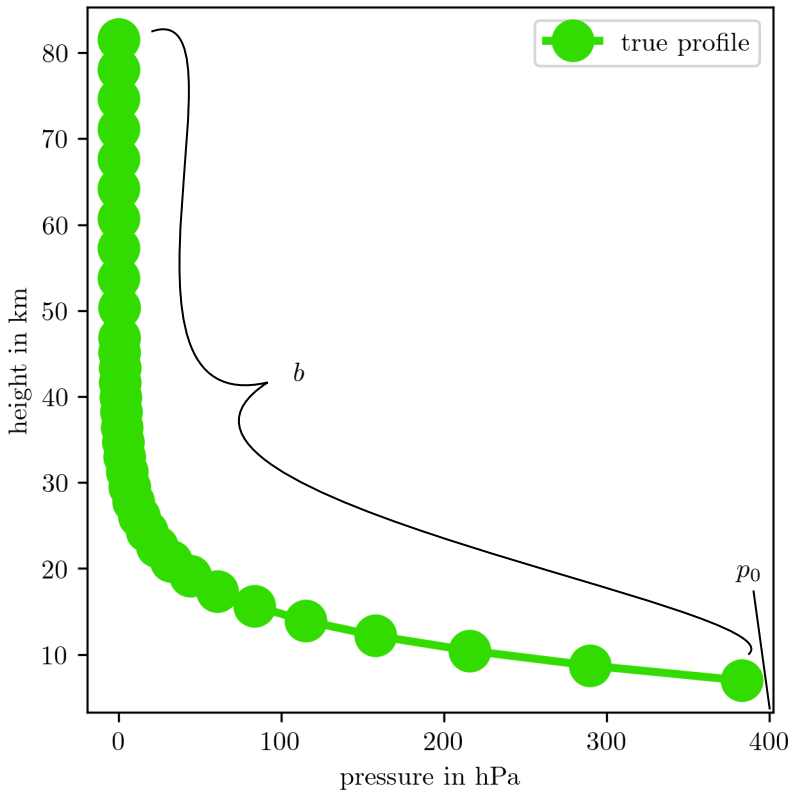


Figure 15: True pressure profile including hyper-parameters as in Eq. 22.

7.2 Linear Forward Model

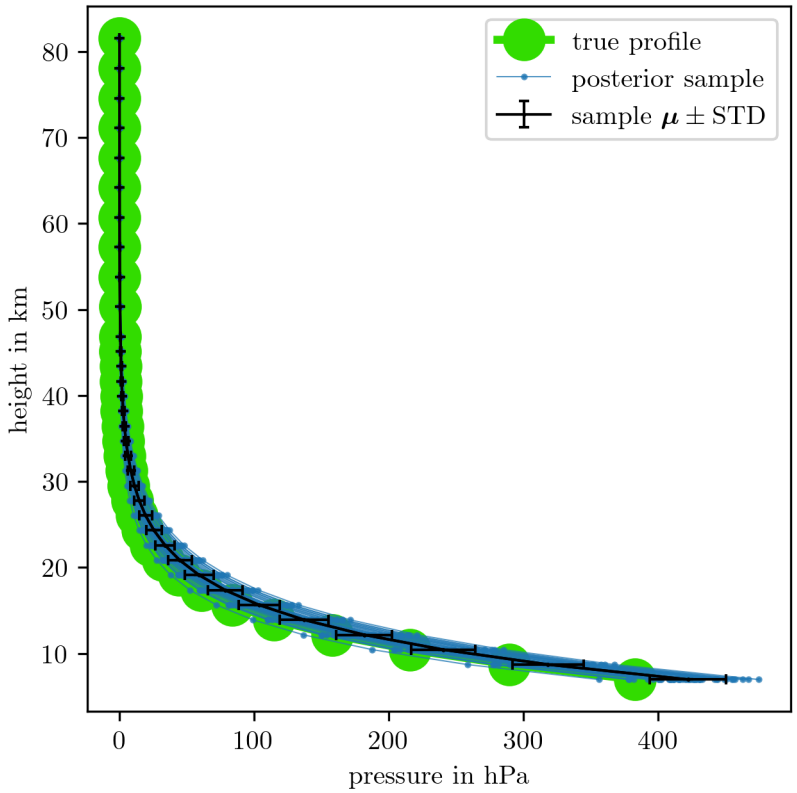


Figure 16: 50 posterior pressure profiles from marginal posterior samples based on the linearised forward model via Eq. 22.

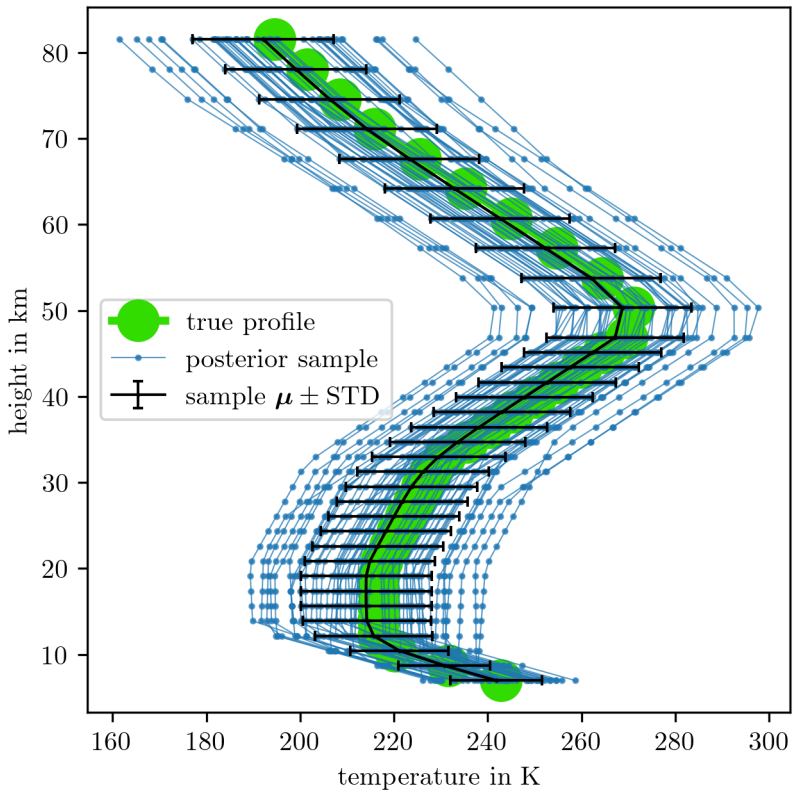


Figure 17: 50 posterior pressure profiles from marginal posterior samples based on the linearised forward model via Eq. 18.

7.3 Approximated Forward Model

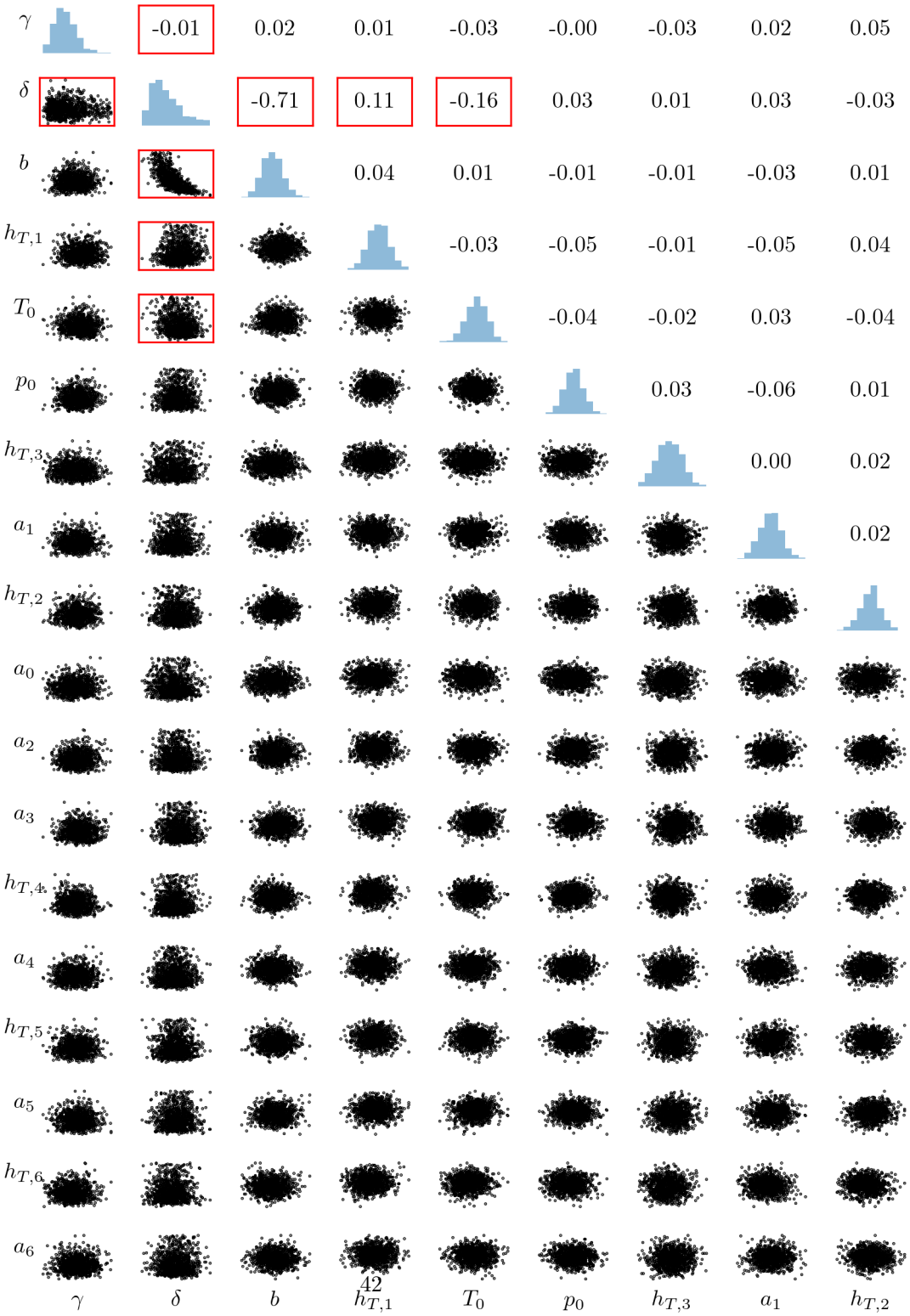
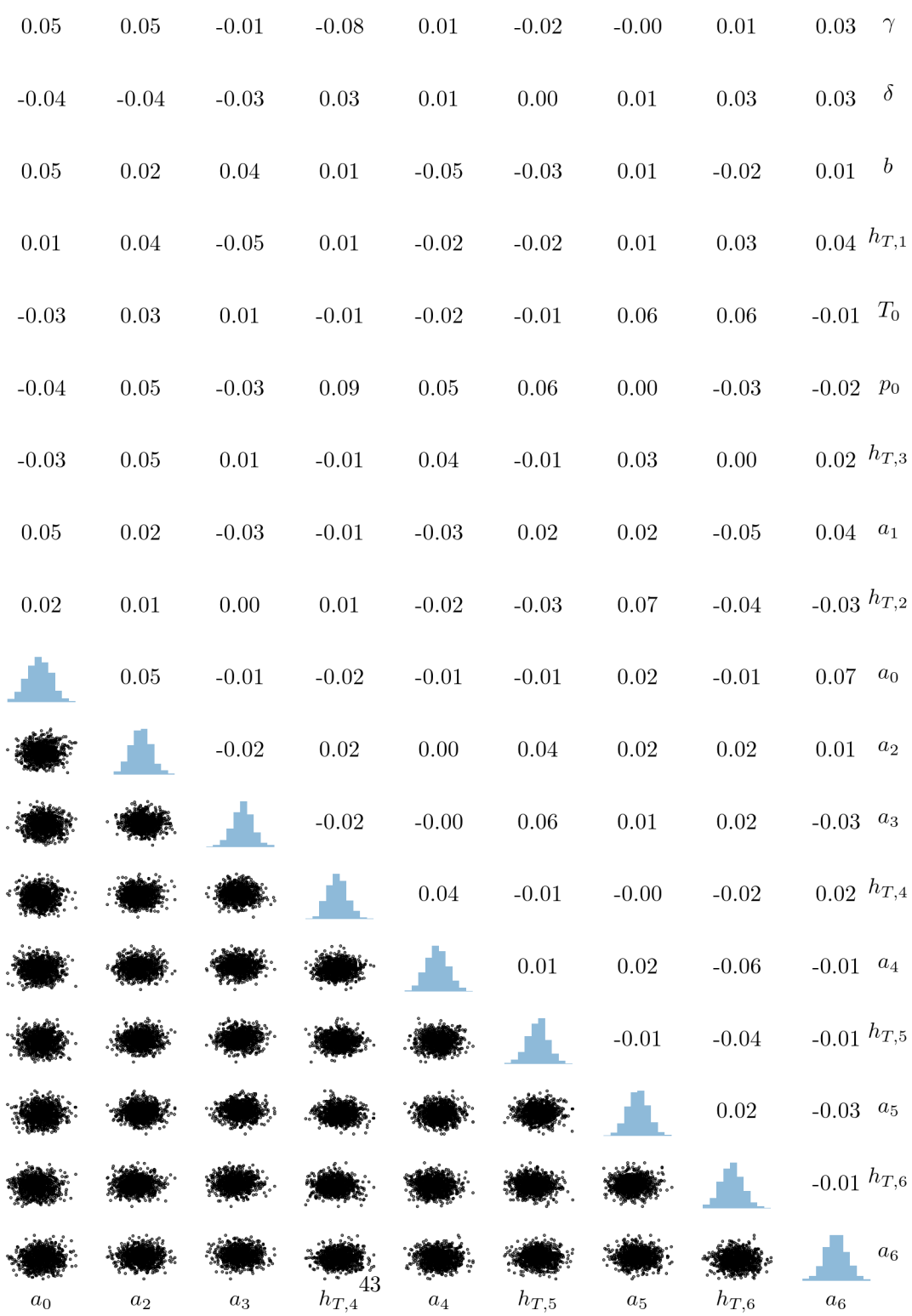


Figure 18: Plot of 1000 independent samples from the marginal posterior $\pi(\boldsymbol{\Phi}, \delta, \gamma|\boldsymbol{y})$ based on the approximated forward model via the SIRT-MH scheme. We plot the Pearson correlation coefficient ranging from -1 to 1 for each hyper-parameter pair.



Correlation plot of samples from the marginal posterior $\pi(\boldsymbol{\Phi}, \delta, \gamma | \boldsymbol{y})$ based on the approximated forward model via the SIRT-MH scheme.