Efficient inference in a weakly non-linear Bayesian inverse Problem using an affine approximation

Lennart Golks and Colin Fox

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 ozone concentrations as well as some 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 (MTC) scheme [8] enables efficient inference.

1 Introduction

There are satellites, which orbit around the earth at a height of around 500km above the ground, 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 [29] and the Michelson Interferometer for Passive Atmospheric Sounding (MIPAS) on ESA's Envisat [19].

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 methods use regularisation methods to retrieve trace gas concentration from measurements [21, 14, 16]. These methods do not include hyper-parameters, such as noise covariance, and produce biased results [9].

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. 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 [8]. This gives a low-dimensional marginal posterior probability distribution over the hyper-parameters and a high-dimensional conditional posterior probability for the ozone parameter. We show that evaluating the marginal posterior on a grid gives the conditional posterior mean at no additional cost. To quantify the conditional posterior variance, samples from the conditional posterior are drawn via the randomise-then-optimise (RTO) scheme [1].

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 unbiased mean and variance of the posterior ozone profile.

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.

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 represntation). Within a Bayesian approach all unknown hyper-parameters and parameters are treated as random variables [12, Chapter 3].

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

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

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

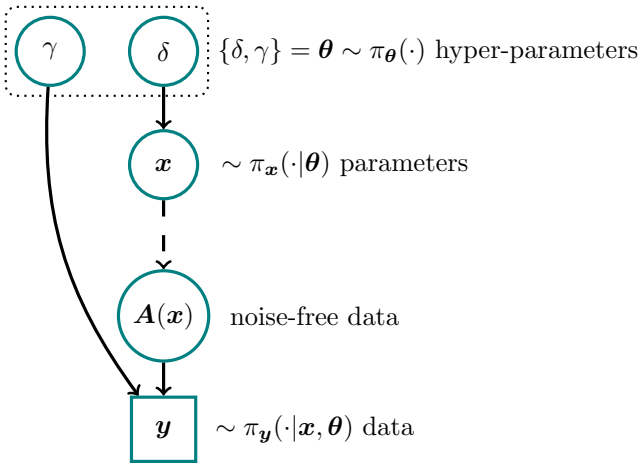

Figure 1: A directed acyclic graph (DAG) for an inverse problem visualises statistical dependencies as solid line arrows and deterministic dependencies as dotted arrows. The hyper-parameters $\boldsymbol{\theta}$ are distributed as ($\sim$) the hyper-prior distribution $\pi(\boldsymbol{\theta})$. The 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 $\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$.

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 [22].

For models where the full conditional over the latent field has a known form, 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-then-conditional (MTC) method [8], 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 [8, Lemma 2] is relatively cheap. Applying the law of total expectation [5], Eq. (4) becomes

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

$$= \int \mathbb{E}_{\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y}}\left[h(\boldsymbol{x})\right]\, \pi(\boldsymbol{\theta}|\boldsymbol{y})\ \mathrm{d}\boldsymbol{\theta} \tag{8}$$

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

In the case of a linear-Gaussian hierarchical Bayesian model, as in [8], 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 [8]). 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 [8] 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 [18] 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.

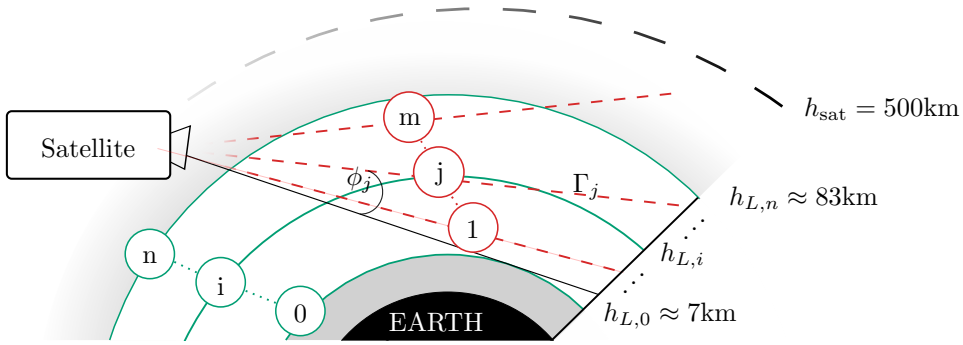

Figure 2: Schematic of measurement and analysis geometry, not to scale. The stationary satellite, at a constant height h_{sat} above Earth, takes m measurements along its line-of-sight defined by the line Γ_j. Each measurement has a pointing angle ϕ_j and a tangent height h_{ℓ_j}, $j = 1, 2, \ldots, m$ 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) [18]

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

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

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 $p(r)/(k_{\text{B}} T(r))$ 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

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

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 cm^{-1} of the targeted transition and the second radiation constant $c_2 \coloneqq hc/k_{\text{B}} \approx 1.44\text{cmK}$ are provided by the HITRAN database [10]. The total internal partition function is given as

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

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 [25]), 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 [23, Chapter 1], and for a more comprehensive model, we refer to [17].

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 $L(\nu, T_{\text{ref}})$ is neglected. Normally, this is modelled as the convolution of the normalised Lorentz profile (collisional/pressure broadening) and the normalised Doppler profile (thermal broadening) [18]. 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 [18].

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 [24] 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. 8, which seems to be a typical nighttime profile [13]. For more information on the processes within the atmosphere for ozone, we refer to [13].

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 $R^* = 8.31432 \times 10^{-3}$Nm/kmol/K and the mean molecular weight of the air is $M = 28.97$kg/kmol, as in [27]. 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 [4, 20]), for $i = 1, \ldots, n$. At sea level $h = 0$km the mean pressure is $p_0 = 1013.25$hPa and the mean temperature is $T_0 = 288.15$K [27]. 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 $r_0 \approx 6356\,\text{km}$, the gravitation at sea level is $g_0 \approx 9.81\text{m/s}^2$. For a ground truth temperature profile we follow [27] and 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 [27]. This function describes the mean temperature in the atmosphere with various height-depending gradients according to the different atmospheric layers, as displayed in Fig. 3. 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 [15, 29]. The corresponding wave number is $\nu = 7.86\text{cm}^{-1}$. The absorption constant $k(\nu, T)$ is calculated as in Eq. 11,

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.

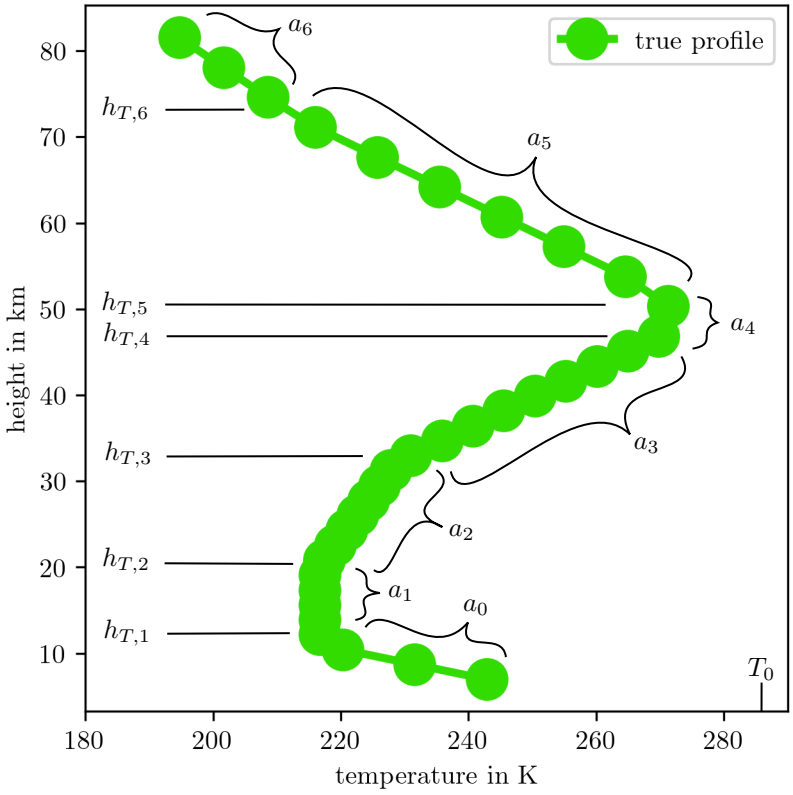

Figure 3: True temperature profile including hyper-parameters as in Eq. 18.

following the high-resolution transmission (HITRAN) database [10]. 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 $h_{\text{sat}} = 500\text{km}$ (see Fig. 2). 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

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

Above $\approx 68\text{km}$ the data is noise dominated (see Fig. 5), hence no measurement are taken in larger 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 $\boldsymbol{A}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{T}) \in \mathbb{R}^m$ for the set of m noise-free measurements. Here, each entry A_j of $\boldsymbol{A}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{T}) \in \mathbb{R}^m$ 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 $\boldsymbol{A}(\boldsymbol{x}) \coloneqq \boldsymbol{A}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{T})$. The simulated data vector

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

includes an additive identically-distributed Gaussian noise vector $\boldsymbol{\eta} \sim \mathcal{N}(0, \gamma^{-1}\boldsymbol{I})$. 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 [9], 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 $\boldsymbol{A}_L \in \mathbb{R}^{m \times n}$. The integral in Eq. (10) is evaluated using the trapezoidal rule and enables matrix-vector multiplication $\boldsymbol{A}_L\boldsymbol{x}$ to compute noise-free linear data. Since neglecting the absorption changes the measurements only slightly (about 1%, see Sec. 5.3.1), 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.

This weakly non-linear forward model may be approximated with an affine map $\boldsymbol{M}$, so that $\boldsymbol{A}(\boldsymbol{x}) \approx \boldsymbol{M}\boldsymbol{A}_L\boldsymbol{x}$. The affine map $\boldsymbol{M} = \boldsymbol{I} + \epsilon$ captures the linear forward model plus a small change ϵ. Note, because of this, we do not necessarily have to find a fixed point to approximate around. In the case of a Taylor series (of first order), one may approximate around the posterior ozone mean, see Eq. 46. The affine map (or the small change ϵ) may be calculated with a linear solver or a least squares fit.

4 Linear-Gaussian Hierarchical Model

First this inverse problem is treated as a linear inverse problem by neglecting the absorption term in the RTE (see Eq. 10), so that the linear forward model matrix is defined as $\boldsymbol{A} \coloneqq \boldsymbol{M}\boldsymbol{A}_L$ with $\boldsymbol{M} = \boldsymbol{I}$. The reader is guided through

the process of setting up a hierarchical Bayesian framework and establishing a choice of prior distributions. Applying the MTC scheme, we explicitly formulate the respective posterior distributions. A directed acyclic graph (DAG) is used to visualise conditional dependencies between hyper-parameters and the parameter (see Fig. 4), 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.

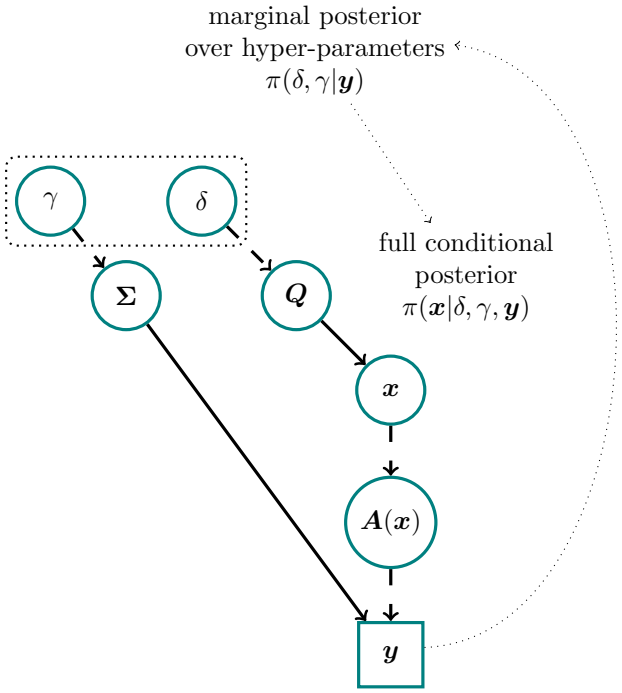

Figure 4: DAG for visualisation of the hierarchical modelling process and the conditional dependency between the parameter and the hyper-parameters. The hyper-parameter γ deterministically (dotted line) sets the noise covariance $\boldsymbol{\Sigma} = \gamma^{-1}\boldsymbol{I}$, which then describes the random (solid line) noise vector $\boldsymbol{\eta} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$. The hyper-parameter δ determines (dotted line) the prior precision matrix $\boldsymbol{Q} = \delta\boldsymbol{L}$ for the normally distributed (solid line) prior $\boldsymbol{x}|\delta \sim \mathcal{N}(0, \boldsymbol{Q}^{-1})$, where $\boldsymbol{L}$ is a graph Laplacian, see Eq. 22. The hyper-prior distributions (solid line) $\pi(\delta, \gamma)$ are defined by $\boldsymbol{\theta}_\gamma$ and $\boldsymbol{\theta}_\delta$. Through the forward model we generate noise-free data $\boldsymbol{A}(\boldsymbol{x})$ and the obeserved data set $\boldsymbol{y}$ includes some added noise $\boldsymbol{\eta}$. Within the MTC scheme, we evaluate the marginal posterior over the hyper-parameters $\pi(\gamma, \delta|\boldsymbol{y})$ first and then the full conditional posterior $\pi(\boldsymbol{x}|\delta, \gamma, \boldsymbol{y})$. This breaks the correlation structure of δ and γ, and $\boldsymbol{x}$, and allows us to evaluate the marginal posterior independently of $\boldsymbol{x}$.

The distributions that define the hierarchical Bayesian model are:

$$\boldsymbol{y}|\boldsymbol{x},\delta,\gamma \sim \mathcal{N}(\boldsymbol{A}\,\boldsymbol{x},\gamma^{-1}\boldsymbol{I}) \tag{21a}$$

$$\boldsymbol{x}|\delta \sim \mathcal{N}(\boldsymbol{0},(\delta\boldsymbol{L})^{-1}) \tag{21b}$$

$$\delta \sim \mathcal{T}(\alpha_\delta,\beta_\delta) \tag{21c}$$

$$\gamma \sim \mathcal{T}(\alpha_\gamma,\beta_\gamma)\,. \tag{21d}$$

Assuming Gaussian noise $\boldsymbol{\eta} \sim \mathcal{N}(0,\gamma^{-1}\boldsymbol{I})$, the likelihood function is a normal distribution with mean $\boldsymbol{A}\boldsymbol{x}$ and covariance matrix $\gamma^{-1}\boldsymbol{I}$. We define a normal prior-distribution $\pi(\boldsymbol{x}|\delta)$ with zero mean and precision matrix $\delta\boldsymbol{L}$, where δ is a smoothness hyper-parameter and $\boldsymbol{L}$ is a discrete approximation to the second derivate operator (see Eq. 22). Here the hyper-prior distributions $\pi(\delta)$ and $\pi(\gamma)$ are Gamma distributions with shape α and rate β.

We can visualise this hierarchical structure and the conditional dependencies between hyper-parameters and parameters through a DAG, as in Fig. 4. The hyper-parameter γ sets the noise covariance deterministically (dotted line), but is itself statistically (solid line) defined by the hyper-prior distribution $\pi(\gamma)$. This is a Gamma distribution, where $\boldsymbol{\theta}_\gamma$ determines the shape and rate of $\pi(\gamma)$. Similarly $\boldsymbol{\theta}_\delta$ defines the Gamma distribution $\pi(\delta)$. Through the linear forward model the space Ω is determined by all measurable noise-free data sets $\boldsymbol{A}\boldsymbol{x}$. From that space we observe (square box) a data set $\boldsymbol{y}$ including some additive noise $\boldsymbol{\eta}$.

Given that data, we "reverse the arrows" to determine the posterior distribution $\pi(\boldsymbol{x},\delta,\gamma|\boldsymbol{y})$ over the parameter $\boldsymbol{x}$ and the hyper-parameters δ and γ. Usually, due to underlying correlation structures between the parameter and the hyper-parameters, evaluating this posterior poses a significant challenge. The MTC scheme breaks this correlation and provides the marginal posterior $\pi(\delta,\gamma|\boldsymbol{y})$ first and then the full conditional posterior $\pi(\boldsymbol{x}|\delta,\gamma,\boldsymbol{y})$.

4.0.1 Prior Modelling

Completing this Bayesian framework one has to define prior distributions over the hyper-parameters and parameters. Ideally, we define the prior distributions as uninformative as possible, and include functional dependencies and physical properties.

By choosing a normally distributed prior $\pi(\boldsymbol{x}|\delta)$ 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

$$\delta\boldsymbol{L} = \delta \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix} \tag{22}$$

which is a discrete approximation to the second derivative operator with Dirichlet boundary condition and defines a 1-dimensional Graph Laplacian as in [28,

8]. 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.

For δ and γ 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 [8]. Because of those Gamma distributions, $\pi(\gamma|\lambda, \boldsymbol{y}) \sim \mathcal{T}(\cdot)$ is a Gamma distribution with $\lambda = \delta/\gamma$ and easy to sample from.

4.1 Posterior Distribution

As explained in Sec. 2.1, we factorise the posterior

$$\pi(\boldsymbol{x}, \delta, \gamma|\boldsymbol{y}) \propto \pi(\boldsymbol{y}|\boldsymbol{x}, \delta, \gamma)\pi(\boldsymbol{x}, \delta, \gamma) \tag{23}$$

into

$$\pi(\boldsymbol{x}, \delta, \gamma|\boldsymbol{y}) = \pi(\boldsymbol{x}|\delta, \gamma, \boldsymbol{y})\pi(\delta, \gamma|\boldsymbol{y}) \tag{24}$$

the marginal posterior $\pi(\delta, \gamma|\boldsymbol{y})$ and full conditional posterior $\pi(\boldsymbol{x}|\delta, \gamma, \boldsymbol{y})$ (see Eq. 5). As discussed in [8], 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 first and then the full conditional posterior.

4.1.1 Marginal Posterior

For the hierarchical model specified in Eq. 21a to Eq. 21d, the marginal posterior distribution over the hyper-parameters is given by

$$\pi(\lambda, \gamma|\boldsymbol{y}) \propto \lambda^{n/2+\alpha_\delta-1} \gamma^{m/2+\alpha_\delta+\alpha_\gamma-1} \exp\left\{-\frac{1}{2}g(\lambda) - \frac{\gamma}{2}f(\lambda) - \beta_\delta\lambda\gamma - \beta_\gamma\gamma\right\}, \tag{25}$$

with the regularisation parameter $\lambda = \delta/\gamma$, and

$$f(\lambda) = \boldsymbol{y}^T\boldsymbol{y} - (\boldsymbol{A}^T\boldsymbol{y})^T(\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{L})^{-1}\boldsymbol{A}^T\boldsymbol{y}\,, \tag{26a}$$

$$\text{and } g(\lambda) = \log\det(\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{L})\,. \tag{26b}$$

Note that when changing variables from $\delta = \lambda\gamma$ to λ the hyper-prior distribution changes to $\pi(\lambda|\gamma) \propto \lambda^{\alpha_\delta - 1}\gamma^{\alpha_\delta} \exp(-\beta_\delta\lambda\gamma)$, due to $\mathrm{d}\delta/\mathrm{d}\lambda = \gamma$. Here, λ is introduces as the regularisation parameter [8]. Because of the chosen Gamma priors the conditional marginal posterior

$$\gamma|\lambda, \boldsymbol{y} \sim \mathcal{T}\left(\frac{m}{2} + \alpha_\delta + \alpha_\gamma, \frac{1}{2}f(\lambda) + \beta_\gamma + \beta_\delta\lambda\right) \tag{27}$$

is a Gamma distribution.

4.1.2 Full Conditional Posterior

As in [26], consider the joint Gaussian distribution

$$\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{A}\boldsymbol{\mu} \end{pmatrix}, \begin{pmatrix} \boldsymbol{Q} + \boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A} & -\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1} \\ \boldsymbol{\Sigma}^{-1}\boldsymbol{A} & \boldsymbol{\Sigma}^{-1} \end{pmatrix}^{-1} \right] \tag{28}$$

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

$$\boldsymbol{x}|\delta,\gamma,\boldsymbol{y} \sim \mathcal{N}\big(\underbrace{(\boldsymbol{A}^T\boldsymbol{A} + \delta/\gamma\boldsymbol{L})^{-1}\boldsymbol{A}^T\boldsymbol{y}}_{\boldsymbol{x}_\lambda}, \underbrace{(\gamma\boldsymbol{A}^T\boldsymbol{A} + \delta\boldsymbol{L})}_{\gamma\boldsymbol{B}_\lambda}{}^{-1}\big)\,, \tag{29}$$

is a normal distribution with $\lambda = \delta/\gamma$ and samples can be drawn via the randomise then optimise (RTO) method (see Sec. 5.2).

Alternatively, the posterior mean

$$\boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{y}} = \int \boldsymbol{x}_\lambda \pi(\lambda|\boldsymbol{y})\,\mathrm{d}\lambda \approx \sum \boldsymbol{x}_{\lambda_i}\pi(\lambda_i|\boldsymbol{y})\,, \tag{30}$$

and posterior covariance

$$\boldsymbol{\Sigma}_{\boldsymbol{x}|\boldsymbol{y}} = \int \gamma^{-1}\pi(\gamma|\boldsymbol{y})\;\mathrm{d}\gamma \int \boldsymbol{B}_\lambda^{-1}\,\pi(\lambda|\boldsymbol{y})\,\mathrm{d}\lambda \approx \sum {\gamma_i}^{-1}\pi(\gamma_i|\boldsymbol{y}) \sum \boldsymbol{B}_{\lambda_i}^{-1}\pi(\lambda_i|\boldsymbol{y}) \tag{31}$$

of $\pi(\boldsymbol{x}|\boldsymbol{y})$ can be computed as weighted expectations over the marginal posterior $\pi(\lambda,\gamma|\boldsymbol{y})$ by quadrature [6, Sec. 2.1] with $\sum \pi(\lambda_i|\boldsymbol{y}) = \sum \pi(\gamma_i|\boldsymbol{y}) = 1$.

5 Results

Given the simulated data the problem is treated as a linear inverse problem, neglecting the absorption in the RTE. The marginal posterior is evaluated on a predefined grid, where the posterior mean is obtained at no additional computational cost. Samples from the conditional posterior are obtained via the RTO method. An affine map approximates the non-linear forward model from the linear forward model. Again, with the approximated forward model, the marginal posterior is calculated on the same predefined grid. Lastly, the mean and the sample-based STD of the posterior ozone profile are provided. Recall that $\boldsymbol{x} \in \mathbb{R}^n$ with $n = 34$ and $\boldsymbol{y} \in \mathbb{R}^m$ with $m = 30$.

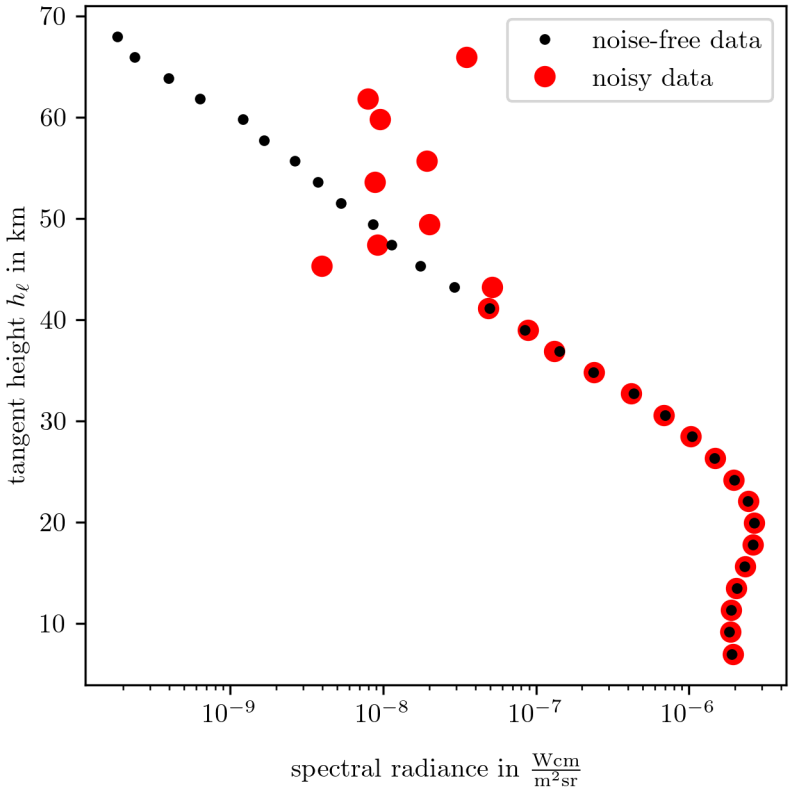

Figure 5: Non-linear noise-free data and non-linear noisy data at different tangent heights. Negative noisy data values are not shown due to logarithmic scaling.

5.1 Marginal Posterior

A $N_{\text{grid}} \times N_{\text{grid}}$ grid for $\lambda = [10^{-5}, 8 \times 10^{-4}]$ and $\gamma = [0.8 \times 10^{15}, 1.2 \times 10^{16}]$ with $N_{\text{grid}} = 25$ grid points is established. The marginal posterior

$$\pi(\lambda, \gamma | \boldsymbol{y}) \propto \lambda^{34/2} \gamma^{30/2+1} \exp\Big\{ -\frac{1}{2} g(\lambda) - \frac{\gamma}{2} f(\lambda) - 10^{-35} \gamma(\lambda - 1) \Big\}, \tag{32}$$

is evaluated on that grid. For each evaluation of the marginal posterior, most of the computational effort lies in $f(\lambda)$ and $g(\lambda)$. For each λ on the grid we do a Cholesky decomposition of $\boldsymbol{B}_\lambda = \boldsymbol{A}^T \boldsymbol{A} + \delta/\gamma \boldsymbol{L} = \boldsymbol{C}^T \boldsymbol{C}$ via the Python function `numpy.linalg.cholesky`. This immediately gives us $g(\lambda) = 2 \sum \log \operatorname{diag}(\boldsymbol{C})$. We calculate $\boldsymbol{x}_\lambda = (\boldsymbol{A}^T \boldsymbol{A} + \lambda \boldsymbol{L})^{-1} \boldsymbol{A}^T \boldsymbol{y}$ via `scipy.linalg.cho_solve` which then gives us $f(\lambda) = \boldsymbol{y}^T \boldsymbol{y} - (\boldsymbol{A}^T \boldsymbol{y})^T \boldsymbol{x}_\lambda$. Then for each λ the cost to compute the marginal posterior $\pi(\lambda, \cdot | \boldsymbol{y})$ for varying $\gamma = [0.8 \times 10^{15}, 1.2 \times 10^{16}]$ is negligible.

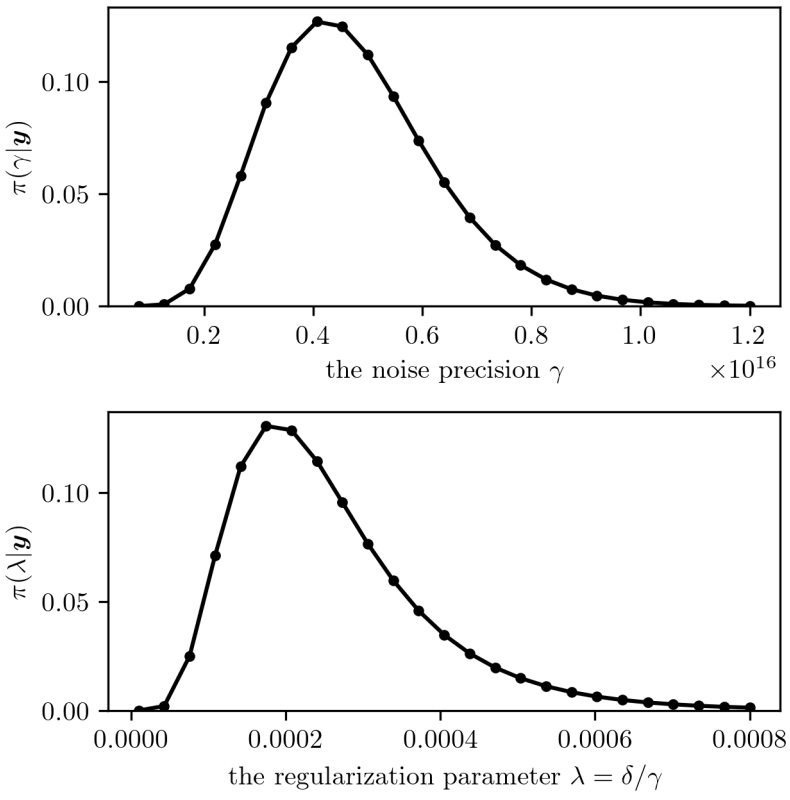

Figure 6: Marginal posterior for linear forward model on a 20×20 grid.

Further, the posterior mean in Eq. 30 is obtained by weighted expatiation

over the predefined grid with $\pi(\lambda|\boldsymbol{y}) = \int \pi(\lambda, \gamma|\boldsymbol{y})\,\mathrm{d}\gamma$ at no additional computational cost. We normalise numerically over the grid so that $\sum \pi(\lambda_i|\boldsymbol{y}) = 1$.

5.1.1 Sampling from Marginal posterior

Independent samples $\lambda^{(k)}, \gamma^{(k)} \sim \pi_{\lambda,\gamma}(\cdot|\boldsymbol{y})$ from the Marginal posterior are needed if one wants to draw an ozone sample from the full conditional posterior $\boldsymbol{x}^{(k)} \sim \pi_{\boldsymbol{x}}(\cdot|\lambda^{(k)}, \gamma^{(k)}, \boldsymbol{y})$.

We sample from the marginal posterior via the inverse Rosenblatt transport (IRT) scheme as in [7]. A seed $\boldsymbol{u}^{(k)} \sim \mathcal{U}(0,1)^d$, where $d = 2$ and $\boldsymbol{u}^{(k)} = \{u_1^{(k)}, u_2^{(k)}\}$, from the uniform distribution is projected onto the hyper-parameter space via the cumulative distribution function (CDF)

$$F(\lambda) = \int_{-\infty}^{\lambda} \pi(\lambda'|\boldsymbol{y})\,\mathrm{d}\lambda'\,, \tag{33}$$

where $\pi(\lambda|\boldsymbol{y}) = \sum_\gamma \pi(\lambda, \gamma|\boldsymbol{y})$ on the predefined grid. The inverse

$$\lambda^{(k)}|\boldsymbol{y} = F^{-1}(u_1^{(k)}) \tag{34}$$

provides a sample from $\pi(\lambda|\boldsymbol{y})$. In order to draw an samples $\gamma^{(k)}|\lambda^{(k)}, \boldsymbol{y}$ we repeat this procedure. $\pi(\gamma|\lambda^{(k)}, \boldsymbol{y})$ (see Eq. 32) is calculated for the every γ value on the grid with fixed $\lambda^{(k)}$. Here we use piecewise linear interpolation via the Python function `numpy.linalg.interp` to obtain the values $f(\lambda^{(k)})$ and $g(\lambda^{(k)})$ (see Fig. 7). Then the CDF

$$F(\gamma) = \int_{-\infty}^{\gamma} \pi(\gamma|\lambda^{(k)}, \boldsymbol{y})\,\mathrm{d}\gamma' \tag{35}$$

is calculated and an independent sample is given as:

$$\gamma^{(k)}|\lambda^{(k)}, \boldsymbol{y} = F^{-1}(u_2^{(k)})\,. \tag{36}$$

Note that alternatively one could draw an independent sample using Eq. 27, which is a Gamma distribution due to the choice of prior. The presented IRT works for arbitrary priors.

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)

$$\tau_{\mathrm{int}} \coloneqq \left(1 + 2\sum_{t=1}^{W} \frac{\Gamma(t)}{\Gamma(0)}\right) \tag{37}$$

as in [8] with an autocorrelation coefficient $\Gamma(t)$. This is twice the value of the IACT in [31, pp. 103-105] and [30, 11], as commonly defined within the physics community. U. Wolff [30] (and the Python implementation by D. Hesse [11]) provide a way to not only calculate the IACT safely but also to quantify the errors of

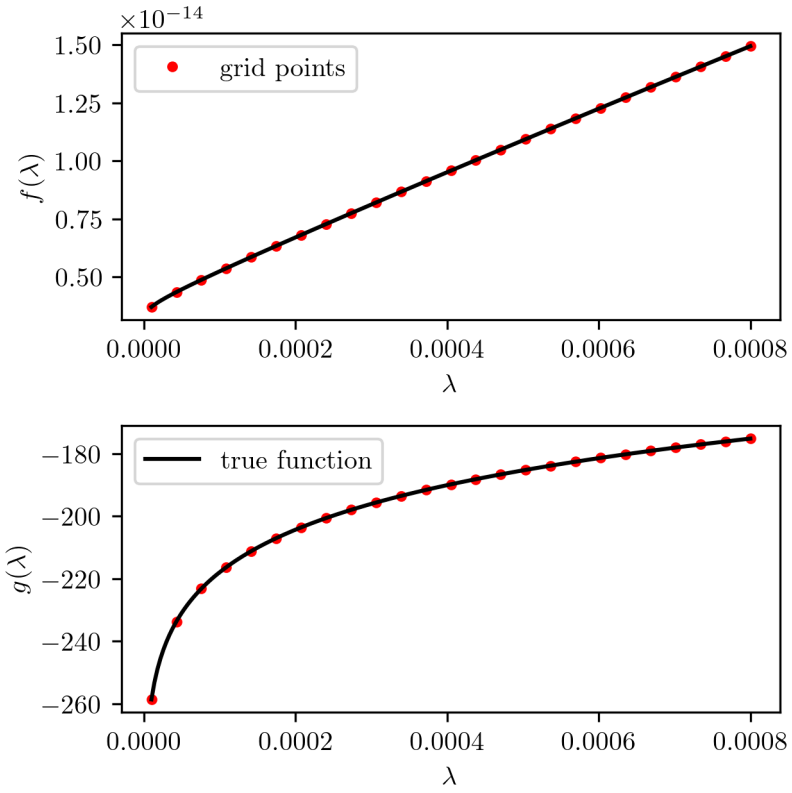

Figure 7: Functions $f(\lambda)$ and $g(\lambda)$ for $\lambda = [10^{-5}, 8 \times 10^{-4}]$, grid points are marked red. True function values between grid points are plotted in black.

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.

5.2 Full Conditional Posterior

To sample an posterior ozone profile from the full conditional $\pi(\boldsymbol{x}|\delta, \gamma, \boldsymbol{y})$ with $\delta = \lambda\gamma$ we use the RTO method, as in [1, 2, 3, 8]. Rewrite the full conditional

posterior as

$$\pi(\boldsymbol{x}|\delta,\gamma,\boldsymbol{y}) \propto \pi(\boldsymbol{y}|\boldsymbol{x},\gamma)\pi(\boldsymbol{x}|\delta) \tag{38}$$

$$\propto \exp\left(-\frac{1}{2}(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y})\right)\exp\left(-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{x})^T\boldsymbol{Q}(\boldsymbol{\mu}-\boldsymbol{x})\right), \tag{39}$$

$$= \exp\left(-\frac{1}{2}\left\|\hat{\boldsymbol{A}}\boldsymbol{x}-\hat{\boldsymbol{y}}\right\|_{L^2}^2\right), \tag{40}$$

where

$$\hat{\boldsymbol{A}} \coloneqq \begin{bmatrix}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{A}\\ \boldsymbol{Q}^{1/2}\end{bmatrix}, \quad \hat{\boldsymbol{y}} \coloneqq \begin{bmatrix}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{y}\\ \boldsymbol{Q}^{1/2}\boldsymbol{\mu}\end{bmatrix}, \tag{41}$$

$\boldsymbol{Q} = \delta\boldsymbol{L}$ 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 $\pi(\boldsymbol{x}|\delta,\gamma,\boldsymbol{y})$ is obtained by minimising the following equation:

$$\boldsymbol{x}^{(k)} = \arg\min_{\boldsymbol{x}} \|\hat{\boldsymbol{A}}\boldsymbol{x}-(\hat{\boldsymbol{y}}+\boldsymbol{b})\|_{L^2}^2 \tag{42}$$

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

$$\left(\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A}+\boldsymbol{Q}\right)\boldsymbol{x}^{(k)} = \boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}+\boldsymbol{Q}\boldsymbol{\mu}+\boldsymbol{v}_1+\boldsymbol{v}_2, \tag{43}$$

with $\boldsymbol{v}_1 \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A})$ and $\boldsymbol{v}_2 \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{Q}^{-1})$, representing independent Gaussian random variables [1, 8].

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

$$\boldsymbol{x}^{(k)} = \Big(\underbrace{\gamma^{(k)}\boldsymbol{A}^T\boldsymbol{A}+\delta^{(k)}\boldsymbol{L}}_{\boldsymbol{B}^{(k)}}\Big)^{-1}\left(\gamma^{(k)}\boldsymbol{A}^T\boldsymbol{y}+\sqrt{\gamma^{(k)}}\boldsymbol{A}^T\tilde{\boldsymbol{v}}_1+\sqrt{\delta^{(k)}}\boldsymbol{L}^{1/2}\tilde{\boldsymbol{v}}_2\right) \tag{44}$$

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 $\boldsymbol{L}$ [1]. 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 `numpy.linalg.cholesky` and `scipy.linalg.cho_solve` is used to solve for $\boldsymbol{x}^{(k)}$. We draw $m-1=29$ posterior samples in ≈ 0.003. The full conditional posterior samples are plotted in Fig. 8 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 $\pi(\boldsymbol{x}|\delta)$. Those full conditional ozone samples are used to find an affine map.

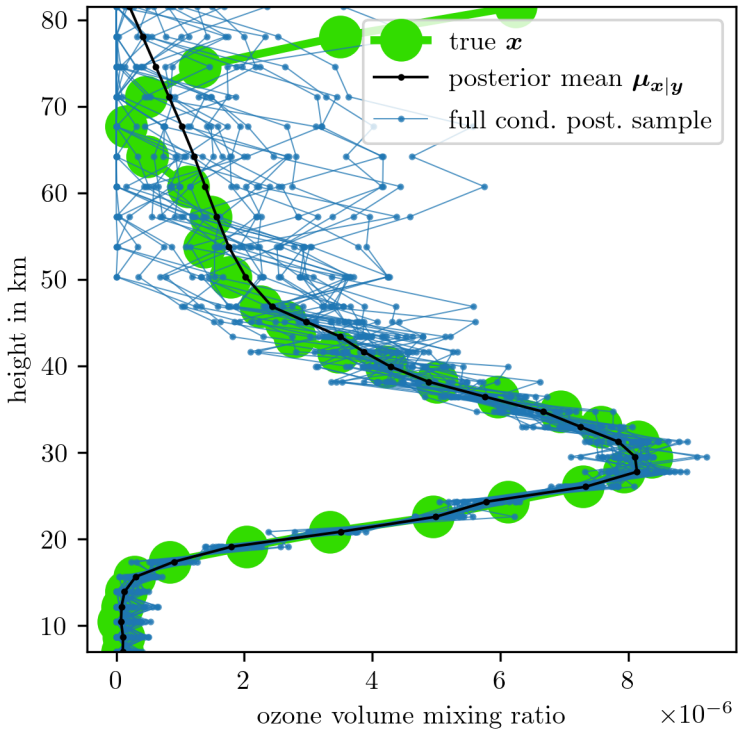

Figure 8: 29 Ozone samples via RTO method and posterior ozone mean via quadrature (see Eq. 30).

5.3 Affine Approximation of the Non-Linear Model

The forward map poses a weakly non-linear forward problem. One could tackle this non-linear inverse problem by fixing the absorption at a previously obtained parameter state and treating this as a linear inverse problem. After each parameter sample the absorption is then iteratively updated. Instead, as in Fig. 9 illustrated, we approximate the non-linear model using an affine map, which is a linear map with a translation, e.g., $\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$. An affine map $\boldsymbol{z} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$ maps a Gaussian $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ onto a Gaussian $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}^T\boldsymbol{\Sigma}\boldsymbol{A})$.

Here we find an affine map $\boldsymbol{M}$ that provides a mapping from the linear forward model $\boldsymbol{A}_L$ to the non-linear model $\boldsymbol{A}(\boldsymbol{x})$ for parameters $\boldsymbol{x}$ near the posterior mean $\boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{y}}$.

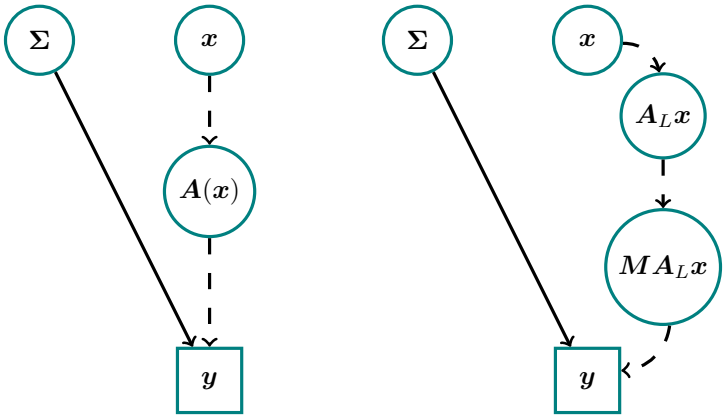

Figure 9: Schematic of how the affine map and the linear forward model approximates the non-linear forward model.

5.3.1 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 $m-1$ samples $\boldsymbol{x}^{(j)} \sim \pi_{\boldsymbol{x}}(\cdot|\delta^{(j)}, \gamma^{(j)}, \boldsymbol{y})$, for $j = 2, \ldots, m$, from the posterior and the posterior mean $\boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{y}}$ generate,

$$\boldsymbol{W} = \begin{bmatrix} | & | & & | & & | \\ \boldsymbol{A}_L\boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{y}} & \boldsymbol{A}_L\boldsymbol{x}^{(2)} & \cdots & \boldsymbol{A}_L\boldsymbol{x}^{(j)} & \cdots & \boldsymbol{A}_L\boldsymbol{x}^{(m)} \\ | & | & & | & & | \end{bmatrix} \in \mathbb{R}^{m \times m}$$

and

$$\boldsymbol{V} = \begin{bmatrix} | & | & & | & & | \\ \boldsymbol{A}(\boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{y}}) & \boldsymbol{A}(\boldsymbol{x}^{(2)}) & \cdots & \boldsymbol{A}(\boldsymbol{x}^{(j)}) & \cdots & \boldsymbol{A}(\boldsymbol{x}^{(m)}) \\ | & | & & | & & | \end{bmatrix} = \begin{bmatrix} — & v_1 & — \\ & \vdots & \\ — & v_j & — \\ & \vdots & \\ — & v_m & — \end{bmatrix} \in \mathbb{R}^{m \times m}\,.$$

Then the non-linear forward model is approximated as

$$\boldsymbol{A}(\boldsymbol{x}) \approx \boldsymbol{M}\boldsymbol{A}_L\boldsymbol{x}\,, \tag{45}$$

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

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

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 $\boldsymbol{V} \in \mathbb{R}^{m \times m}$ is independent of each other as well. For an $\boldsymbol{x} = \boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{y}} + \Delta\boldsymbol{x}$ we rewrite Eq. 45 to

$$\boldsymbol{A}(\boldsymbol{x}) \approx \underbrace{\boldsymbol{M}\boldsymbol{A}_L\boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{y}}}_{=\boldsymbol{A}(\boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{y}})} + \underbrace{\boldsymbol{M}\boldsymbol{A}_L\Delta\boldsymbol{x}}_{=\boldsymbol{A}'(\boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{y}})\Delta\boldsymbol{x}} \tag{46}$$

$$= \underbrace{\boldsymbol{A}'(\boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{y}})\boldsymbol{x}}_{\boldsymbol{A}\boldsymbol{x}} + \underbrace{\boldsymbol{A}(\boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{y}}) - \boldsymbol{A}'(\boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{y}})\boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{y}}}_{\boldsymbol{b}} \tag{47}$$

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

The relative RMS difference $\|\mathrm{vec}(\boldsymbol{M}\boldsymbol{W}) - \mathrm{vec}(\boldsymbol{V})\|_{L^2} / \|\mathrm{vec}(\boldsymbol{M}\boldsymbol{W})\|_{L^2}$ 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 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 $\approx$ 0.01%. 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 $\approx$ 0.1s.

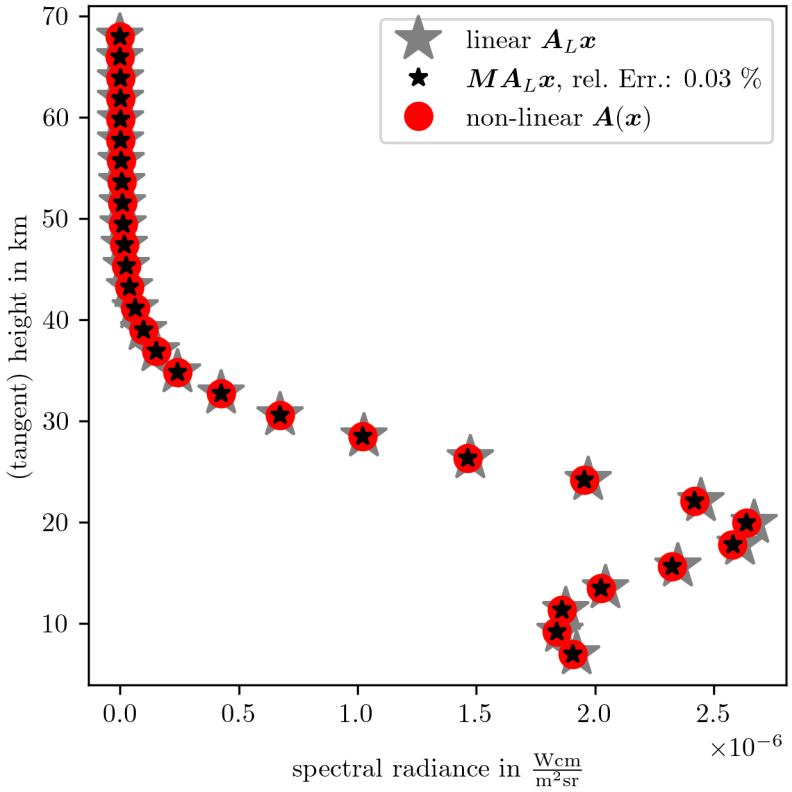

Figure 10: Assessment of how well the affine map $\boldsymbol{M}$ approximates noise-free non-linear data $\boldsymbol{A}(\boldsymbol{x})$ (red circles) from noise-free linear data $\boldsymbol{A}_L\boldsymbol{x}$ (grey stars). The approximated noise-free data (black stars) has a relative RMS error of $\approx 0.01\%$ 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.

5.4 Posterior ozone

The linear forward model that approximates the non-linear forward model is defined as

$$\boldsymbol{A} \coloneqq \boldsymbol{M}\boldsymbol{A}_L \,. \tag{48}$$

We use the exact same grid and setup as in Sec. 5.1 to define the marginal posterior (see Fig. 11). As already mentioned, that gives us the posterior mean

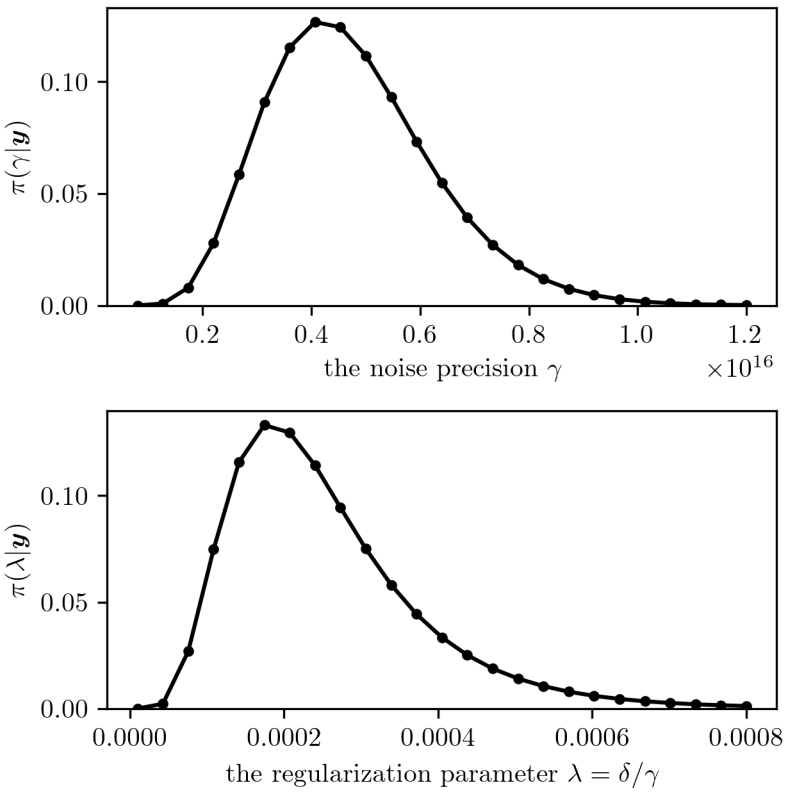

Figure 11: Marginal posterior for approximated forward model on a 20×20 grid.

via weighted expectation for free. The IATCs of the marginal posterior samples are $\tau_{\text{int},\lambda} = 0.97 \pm 0.03$ and $\tau_{\text{int},\gamma} = 0.95 \pm 0.03$ based on a chain with length of 10000. Hence, every second sample is an independent sample from the marginal posterior.

Samples from the conditional posterior are drawn via the RTO method. To estimate the number of samples needed for a good enough estimate of the STD,

we calculate the average Coefficient of Variation ($\overline{\mathrm{CV}}$),

$$\overline{\mathrm{CV}} = \frac{1}{n} \left\| \frac{\sqrt{\mathrm{var}_N(\boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{y}})}}{\boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{y}}} \right\|_1 , \tag{49}$$

recall that $\boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{y}} \in \mathbb{R}^n$ with $n = 34$. The $\overline{\mathrm{CV}}$ gives a measure of how much the STD varies compared to the mean.

If the STD is large compared to the mean the $\overline{\mathrm{CV}}$ is large, if the STD is small compared to the mean the $\overline{\mathrm{CV}}$ is small and a rather accurate estimate of the STD is needed. The sample-based estimate of the variance

$$\mathrm{var}_N(\mu_{\boldsymbol{x}|\boldsymbol{y}}) = \frac{1}{N-1} \sum_{k=1}^{N} (\boldsymbol{x}^{(k)} - \mu_{\boldsymbol{x}|\boldsymbol{y}})^T (\boldsymbol{x}^{(k)} - \mu_{\boldsymbol{x}|\boldsymbol{y}}) \tag{50}$$

is calculated with the posterior mean $\mu_{\boldsymbol{x}|\boldsymbol{y}}$ given by weighted expectations as in Eq. 30. Alternatively, one could evaluate the posterior covariance as a weighted expectation over the marginal posterior grid (see Eq. 31).

The $\overline{\mathrm{CV}}$ is calculated for each number of samples ranging from 1 to 10^3 and plotted in Fig. 12. Fig. 12 shows that the $\overline{\mathrm{CV}}$ is rather large, indicating that

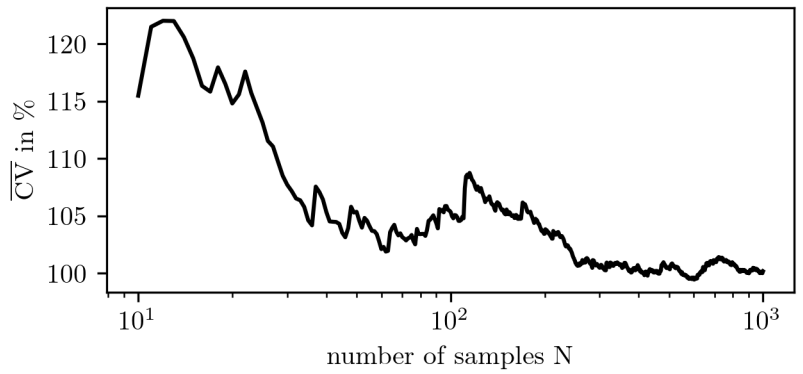

Figure 12: Average Coefficient of Variation in % as calculated in Eq. 49.

the STD is large compared to the posterior mean. Further, we conclude that 50 full conditional posterior ozone samples are enough for a sufficient estimate of the STD. Approximating the marginal posterior on a 20×20 grid and taking 50 independent full posterior samples of ozone takes ≈ 0.008s. The sample-based STD with the posterior mean (via weighted expectation) is plotted in Fig. 13. Note that the posterior ozone profile does not capture the second ozone peak at around 80km.

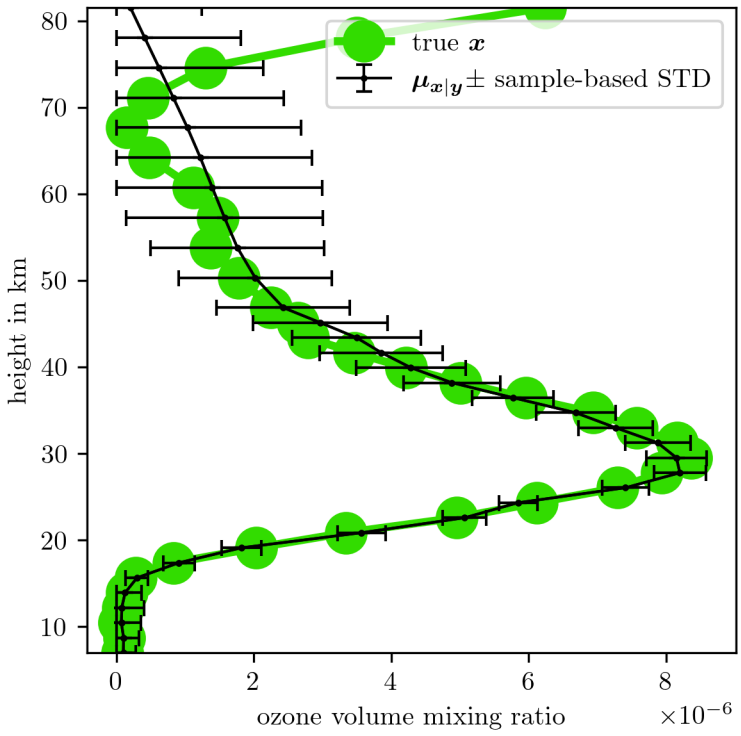

Figure 13: Posterior ozone mean via quadrature (see Eq. 30) and STD based on 20 samples via the RTO method.

6 Conclusion

- 20 solves of x_λ for marginal posteirior and full posterior mean (see Eq. 29 and Eq. 30)
- 29 samples of full conditional posterior are enough for finding an affine map, this includes 29 solves of "randomised" x_λ via the RTO method (see Eq. 44)
- 50 samples of full conditional posterior are enough for the STD estimate of posterior ozone, this includes 50 solves of "randomised" x_λ via the RTO method (see Eq. 44)

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] Dick, J., Kuo, F. Y., and Sloan, I. H. "High-dimensional integration: The quasi-Monte Carlo way". In: *Acta Numerica* 22 (2013), pp. 133–288.

[7] 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.

[8] 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.

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

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

[11] 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].

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

[13] 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).

[14] 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.

[15] 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).

[16] 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.

[17] 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.

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

[19] 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.

[20] Ridolfi, M. et al. "Optimized forward model and retrieval scheme for MIPAS near-real-time dataprocessing". In: *Applied Optics* 39.8 (2000), pp. 1323–1340.

[21] 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.

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

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

[24] 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.

[25] 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.

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

[27] *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.

[28] 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.

[29] 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.

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

[31] 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.