

METHODS PAPER  

Aerosol optical depth disaggregation: toward global aerosol vertical profiles

Shahine Bouabid¹ , Duncan Watson-Parris² , Sofija Stefanović³, Athanasios Nenes⁴ and Dino Sejdinovic⁵

¹Department of Statistics, University of Oxford, Oxford, UK

²Scripps Institution of Oceanography and Halicioğlu Data Science Institute, University of California San Diego, San Diego, CA, USA

³Department of Earth Sciences, University of Cambridge, Cambridge, UK

⁴Laboratory of Atmospheric Processes and their Impacts, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

⁵School of Computer and Mathematical Sciences & AIML, University of Adelaide, Adelaide, SA, Australia

Corresponding author: Shahine Bouabid; Email: shahine.bouabid@stats.ox.ac.uk

Received: 08 May 2022; **Revised:** 11 January 2024; **Accepted:** 10 April 2024



Keywords: aerosol extinction profiles; AOD; ECHAM-HAM; Gaussian processes; vertical disaggregation

Abstract

Aerosol-cloud interactions constitute the largest source of uncertainty in assessments of anthropogenic climate change. This uncertainty arises in part from the difficulty in measuring the vertical distributions of aerosols, and only sporadic vertically resolved observations are available. We often have to settle for less informative *vertically aggregated* proxies such as aerosol optical depth (AOD). In this work, we develop a framework for the vertical disaggregation of AOD into extinction profiles, that is, the measure of light extinction throughout an atmospheric column, using readily available vertically resolved meteorological predictors such as temperature, pressure, or relative humidity. Using Bayesian nonparametric modeling, we devise a simple Gaussian process prior over aerosol vertical profiles and update it with AOD observations to infer a distribution over vertical extinction profiles. To validate our approach, we use ECHAM-HAM aerosol-climate model data which offers self-consistent simulations of meteorological covariates, AOD, and extinction profiles. Our results show that, while very simple, our model is able to reconstruct realistic extinction profiles with well-calibrated uncertainty, outperforming by an order of magnitude the idealized baseline which is typically used in satellite AOD retrieval algorithms. In particular, the model demonstrates a faithful reconstruction of extinction patterns arising from aerosol water uptake in the boundary layer. Observations however suggest that other extinction patterns, due to aerosol mass concentration, particle size, and radiative properties, might be more challenging to capture and require additional vertically resolved predictors.

Impact Statement

Aerosol-cloud interactions (ACIs) represent the largest uncertainty in assessments of global warming, with uncertainty bounds that could offset global warming or double its effects. This uncertainty arises in part from the inability to observe aerosol amounts at an appropriate resolution. Instead aerosol optical depth (AOD)—observed globally through satellite products—is commonly used as a two-dimensional proxy for aerosols. Yet, obtaining a global estimate of aerosol vertical distribution in the atmosphere would be more informative and

  This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

© The Author(s), 2024. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

help better constrain ACIs uncertainty. Here, we show that using AOD and readily available vertically resolved meteorological predictors (temperature, pressure, relative humidity, updraft), a simple probabilistic model can already yield realistic vertical extinction profiles together with appropriate uncertainty quantification. This highlights that such simple modeling can benefit satellite products, leading to more accurate priors over aerosol vertical profiles.

1. Introduction

Aerosols are microscopic particles ($< 5 \mu\text{m}$) suspended in the atmosphere. They can come from natural sources (e.g., dust, sea salt) or be emitted by human activity (e.g., black carbon).

They influence the Earth's energy budget with a negative radiative forcing that counteracts global warming from anthropogenic greenhouse gases emissions. A fraction of this negative forcing stems from aerosols' direct scattering of incoming solar radiation (McCormick and Ludwig, 1967): this is the *direct effect*. A larger fraction of this forcing is due to their modulation of the radiative properties of clouds: this is the *indirect effect*. By acting as cloud condensation nuclei (CCN), additional aerosols can drive up the cloud droplet number while driving down the mean cloud droplet size. The resulting clouds are brighter, larger, and last longer (Twomey, 1977; Albrecht, 1989). They hence reflect more solar radiation and cool the Earth.

Unlike the direct effect which can, in principle, be well constrained (Watson-Parris et al., 2020), the magnitude of the forcing induced by the indirect effect is difficult to estimate. There are two reasons for this: (i) the physical processes underpinning aerosol-cloud interactions (ACIs) are not yet fully understood, which hinders the estimation of present-day forcing; (ii) the present-day forcing must be compared to the forcing at pre-industrial state, which is also particularly challenging to quantify (Carslaw et al., 2013). In fact, observational and model-based studies of ACIs still disagree on the magnitude of this forcing. As a result, ACIs contribute the largest uncertainty in present-day global warming (Masson-Delmotte et al., 2021).

To better estimate present-day forcing we require accurate, global measurements of CCN concentrations to assess radiative properties of clouds (Twomey, 1977; Albrecht, 1989). Unfortunately, measuring CCN concentrations can only be achieved in situ, and while field campaigns have already been undertaken to collect detailed CCN observations, these measurements are spatio-temporally sparse and provide insufficient constraints on the global distribution of aerosols (Andreae, 2009; Spracklen et al., 2011).

For lack of better observations, the aerosol optical depth (AOD) has been widely adopted as a first-order proxy of CCN concentration in an atmospheric column (Nakajima et al., 2001; Andreae, 2009; Clarke and Kapustin, 2010). The AOD is a measure of the extinction of solar radiation through an atmospheric column. It is denoted by τ and defined at a given wavelength, time, latitude, and longitude by

$$\tau = \int_0^H b_{\text{ext}}(h) dh, \quad (1)$$

where b_{ext} is the extinction coefficient¹ and the integral is taken over the height H of an atmospheric column. The AOD is appealing because it is routinely observed on a global scale by satellite products (Remer et al., 2005) which, unlike in situ observations, offer long-term global records.

However, the AOD is a column-integrated quantity and does not provide information on the vertical distribution of aerosols. This is limiting as their vertical distribution strongly influences both the magnitude and even the sign of the forcing induced by the indirect effect. For example, both modeling (Stier, 2016) and observational studies (Painemal et al., 2020) find AOD inadequate for assessing ACIs over vast subtropical ocean areas, which play a key role in determining the radiation balance of the Earth. Yet, in both studies, the vertically resolved aerosol extinction coefficient b_{ext} shows a significantly higher

¹ The sum of contribution from particle-light scattering plus the absorption of light by particles.

correlation with CCN concentrations. Stier (2016) also highlights the importance of determining aerosol vertical distributions to provide stronger constraints on CCN at specific altitudes. In particular, the AOD fails to describe near-surface properties such as the concentration of aerosols in the boundary layer.

In this work, we propose to probe whether AOD observations can be used to constrain a global prior over aerosol vertical distributions. Formally, given an AOD observation τ , we want to reconstruct the corresponding extinction coefficient profile b_{ext} . This amounts to the task of reconstructing three-dimensional (3D) profiles using height-integrated two-dimensional (2D) observations and quantities that are easier to obtain in 3D, such as temperature and relative humidity.

Motivated by the study of cloud vertical structures, this task has been framed in the past as fully supervised learning (Leinonen et al., 2019), that is, assuming observations of groundtruth vertical profiles were available. Collecting vertically resolved observations of aerosols optical properties is also possible, using lidar-based remote sensing instruments (Winker et al., 2013) or ground-based sun-photometers (Holben et al., 1998). While valuable, these observations are however limited by their low spatiotemporal coverage and prone to corruption (e.g., low signal-to-noise ratio, clear-sky requirement). Compiling high-quality observational data of aerosol vertical profiles at a large scale is thus challenging, making fully supervised learning approaches inadequate.

Instead, we propose to draw from spatial disaggregation methodologies that only require observations at the aggregated level. Spatial disaggregation is the task of inferring subgrid details given coarse resolution spatial observations. Postulating an underlying fine-grained spatial field that aggregates into coarse observations, this problem can be framed as weakly supervised learning (Zhou, 2017) with aggregated targets. While existing works (Law et al., 2018; Tanaka et al., 2019; Yousefi et al., 2019; Tanskanen and Longi, 2020; Zhang et al., 2020) have only considered aggregation processes happening on a 2D field, this rationale can be extended to disaggregate quantities along a third dimension—height. Since τ corresponds to the vertical integration of b_{ext} , we propose to frame the reconstruction of aerosol vertical profiles as the vertical disaggregation of AOD observations.

Using Gaussian processes (GPs) (Rasmussen and Williams, 2005), we design a Bayesian model that maps vertically resolved meteorological variables (e.g., pressure, relative humidity) to a probabilistic estimate of the extinction coefficient that integrates into the AOD. The model formulation is simple and makes assumptions explicit, hence granting control and interpretability over predictions while offering built-in uncertainty quantification.

In order to be able to fully validate the proposed methodology, we use ECHAM-HAM global aerosol-climate model simulation data (Stier et al., 2005, 2007; Zhang et al., 2012). While our prime motivation is to reconstruct b_{ext} from satellite observations of AOD, the intricacies of combining measurements from different instruments make it challenging to validate any proposed methodology. On the other hand, ECHAM-HAM is a self-consistent climate model that offers readily available aerosol vertical profiles, and is better suited for model development. We demonstrate our model is able to reconstruct natural patterns that arise in aerosol vertical distribution, in particular in the boundary layer. We show that very simple and readily available meteorological predictors suffice to obtain a good estimation of the extinction coefficient.

The aims of this study are as follows:

- Formulate a probabilistic vertical disaggregation methodology for the task of reconstructing aerosol vertical profiles using readily available vertically resolved covariates.
- Validate the proposed methodology using climate model data—where access to groundtruth extinction coefficient profiles enables quantitative evaluation.

Section 2 outlines the design of the vertical disaggregation model and describes the inference procedure as well as the hyperparameter selection. Section 3 describes the dataset and experimental setup used to validate the model. Section 4 presents the experimental results and Section 5 discusses, while introducing avenues of future research.

2. Model design

In this section, we first outline the design of the prior distribution we place on the extinction coefficient profile using GPs. We then describe the observation model that connects our prior to the observations of the AOD. Finally, we present how the posterior inference over the extinction profiles is conducted.

2.1. Design of the prior

2.1.1. An idealized vertical prior

In passive satellite sensors, AOD retrieval algorithms are used to produce an estimate of the AOD from measurements of top-of-atmosphere reflectance. They are typically based on a forward model of top-of-atmosphere reflectance (Wu et al., 2017). The parameters required to run this forward model are obtained using a look-up-table algorithm, which admits as indices aerosol properties such as the aerosol type, the AOD, and the form of the vertical profile. When the satellite sensor captures a measure of reflectance, the look-up-table can be used to solve the corresponding inverse problem and estimate the AOD, provided the aerosol properties have been predetermined. Consequently, to address the inverse problem, AOD retrieval algorithms need to assume a form for aerosol vertical profile.

The form of the vertical profile is typically idealized, using either a Gaussian profile, or in the simplest case an exponential profile (Wu et al., 2017; Li et al., 2020). The exponential profile takes the form $b_{\text{ext}}(h) \propto e^{-h/L}$, where L is a fixed height scale parameter that is typically taken as the top altitude of the boundary layer (2 km)—although different algorithms may adopt varying values.

Whilst idealized, these profiles capture a key element of aerosol vertical distribution: most CCN lies at low altitude in the boundary layer. This constitutes a form of domain knowledge, which we argue is important to encode within the design of a prior over aerosol vertical profiles. Thus, we propose to use an idealized exponential profile as a structural component of the prior placed over the extinction profile.

2.1.2. Weighting the ideal profile

Clouds' local meteorology influences clouds properties, and hence the sign and magnitude of the effective radiative forcing due to ACIs (Ackerman et al., 2004; Small et al., 2009; Chen et al., 2014; Douglas and L'Ecuyer, 2020)—this source of heterogeneity is at the heart of the uncertainty over how clouds impact climate.

The impact of local meteorology on ACIs can be characterized by the following set of environmental confounders (Köhler, 1936; Twomey, 1974): temperature T , pressure P , relative humidity RH, and vertical velocity or updraft ω . Standing as a proxy for aerosols, the AOD is also impacted by its surrounding meteorology (Christensen et al., 2017; Jesson et al., 2022). These meteorological variables should thus also modulate the extinction coefficient.

A great advantage of these simple meteorological variables is that they (or their proxies) are readily available on multiple pressure levels in reanalysis data. Reanalysis data, such as ERA5 (Muñoz-Sabater et al., 2021), combines observations and model simulations through physical laws to provide the most accurate representation of past and present climate and meteorology. Hence, T , P , RH, and ω can be reliably used as vertically resolved predictors. Finally, note that b_{ext} is also a spatiotemporal field which exhibits smoothness across spatial and temporal dimensions. Such regularity should be included in the modeling.

Let $x|h$ denote a d -dimensional vector resulting from the concatenation of spatiotemporal and meteorological variables for a given altitude h . For example, one can consider

$$x = (t, \text{lat}, \text{lon}, T, P, \text{RH}, \omega), \quad (2)$$

where lat and lon respectively denote latitude and longitude. In the remainder of this paper, we will denote $\mathcal{X} \subseteq \mathbb{R}^d$ the space in which x takes values.

We propose to model the extinction coefficient b_{ext} by weighting an idealized exponential vertical profile with a positive weight function $w: \mathcal{X} \rightarrow (0, +\infty)$. Namely, the chosen prior for the extinction coefficient profile is denoted φ and takes the simple form

$$\varphi(x|h) = w(x|h)e^{-h/L}. \quad (3)$$

This weight function is meant to capture finer details of variability in the extinction profile, putting more mass in regions where meteorological predictors suggest aerosol loading is likely to be higher.

2.1.3. Probabilistic modeling of the weighting function

Whilst the extinction coefficient should indeed be modulated by $x|h$, we expect this relationship to be non-trivial and highly non-linear (Jesson et al., 2022). For this reason, we propose to learn the weighting function w using non-linear (e.g., kernel-based) statistical machine-learning methodologies. Furthermore, we want the prior to reflect our lack of knowledge about the relationship between $x|h$ and $b_{\text{ext}}(h)$. To account for this epistemic uncertainty, we propose a Bayesian formulation of the weighting function.

GPs (Rasmussen and Williams, 2005) are a ubiquitous class of expressive Bayesian priors over real-valued functions. They have been widely used in various nonlinear and nonparametric regression problems in geosciences (Camps-Valls et al., 2016). A GP is fully determined by its mean function and its covariance function. The covariance function—called kernel—is typically user-specified as a positive definite bivariate function on the input data.

We place a GP prior over the weight function. To ensure the weights are strictly positive, we further warp the GP with a positive transform $\psi: \mathbb{R} \rightarrow (0, +\infty)$. Formally, let $m: \mathcal{X} \rightarrow \mathbb{R}$ be a mean function and $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ denote a positive definite kernel, we model the weighting function as

$$w(x|h) = \psi(f(x|h)) \quad \text{where } f \sim \text{GP}(m, k). \quad (4)$$

The Bayesian prior defined by f represents a latent variable that maps through the positive link function ψ onto the weight function. A simple choice for ψ is the exponential function, which as we will see in Section 3 is a mathematically convenient choice.

Whilst $\psi \circ f$ describes an expressive probability distribution over nonlinear positive functions, it remains interpretable. The choice of kernel specifies how covariance is encoded, that is, $\text{Cov}(f(x), f(x')) = k(x, x')$ and controls the functional class the GP belongs to. For example, the Matérn class of covariance functions offers control over the functional smoothness of the GP (Stein, 1999; Rasmussen and Williams, 2005).

2.2. Design of the observation models

In this section, we establish a connection between the extinction coefficient prior $\varphi(x|h)$ (constructed in Section 2.1) and the observed values of extinction coefficient and AOD. To simplify notations, we focus on a single air column with a height of H , where we have observations of the extinction coefficient $b_{\text{ext}}(h)$ and the AOD τ .

Ideally, we would like to achieve a perfect match between $b_{\text{ext}}(h)$ and $\varphi(x|h)$, as well as between τ and $\int_0^H \varphi(x|h)$. Unfortunately, this is unrealistic because observations are prone to measurement error, which must critically be accounted for. To address this, we formulate probabilistic observation models for b_{ext} and τ in what follows.

2.2.1. Relevance of the log-normal distribution

The log-normal distribution has been reported to provide a good fit to the AOD in studies focusing on locations in North America and Europe (O'Neill et al., 2000). This distribution is particularly suitable because the AOD is strictly positive and exhibits a right-skewed distribution concentrated around small values (≈ 0.14).

The log-normal distribution is a right-skewed continuous probability distribution with support $(0, +\infty)$. It is specified by two parameters: a location parameter $\mu \in \mathbb{R}$, and a scale parameter $\sigma > 0$. In particular, if $Z \sim \mathcal{N}(0, 1)$ is a standard normal random variable, then $e^{\mu + \sigma Z}$ follows a log-normal distribution with location μ and scale σ , which we denote $\mathcal{LN}(\mu, \sigma)$.

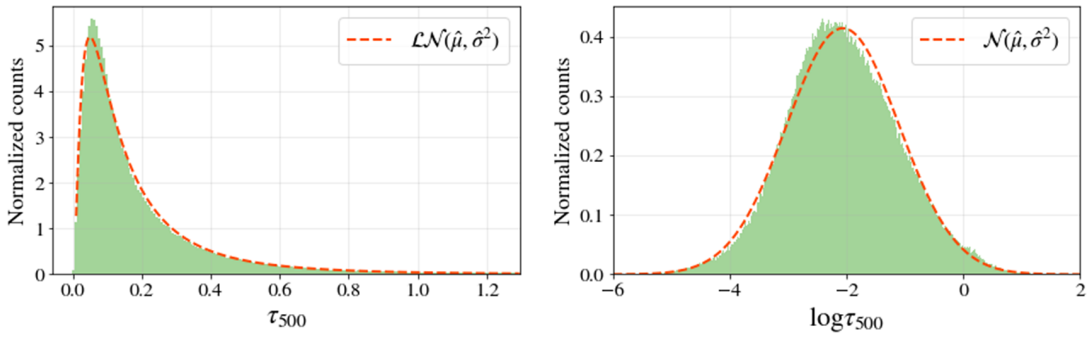


Figure 1. Left: Log-normal density (red) fitted with maximum likelihood estimates to the empirical distribution of AERONET AOD at 500 nm (τ_{500}) from 1315 stations between 1993 and 2021 (green). Right: logspace plot of the left panel. It demonstrates a sound fit for the normal distribution in the logspace, albeit with a slight right skew; $\hat{\mu} = -2.06$, $\hat{\sigma} = 0.96$.

It is often preferable to express the log-normal distribution as a function of its mean rather than its location, as it is more interpretable and offers a convenient parametrization for including higher-order model parameters. Formally, let $\eta = \mathbb{E}[\mathcal{LN}(\mu, \sigma)]$ denote the expectation of the distribution and suppose the scale σ is known. The η can be expressed using μ and σ as $\eta = e^{\mu + \sigma^2/2} \Leftrightarrow \mu = \log \eta - \frac{\sigma^2}{2}$, which yields the mean-reparametrized formulation of the log-normal distribution

$$\mathcal{LN}(\mu, \sigma) = \mathcal{LN}\left(\log \eta - \frac{\sigma^2}{2}, \sigma\right). \tag{5}$$

We collect AERONET sun-photometers AOD measurements from 1315 stations between 1993 and 2021 (Holben et al., 1998), and shown in Figure 1 that the empirical marginal distribution of AERONET observations can indeed be appropriately fitted using a log-normal density. Whilst Figure 1 only depicts a marginal distribution for the AOD, we draw motivation from this to formulate a log-normal observation model for b_{ext} and τ conditional on $\varphi(x|h)$.

2.2.2. An observation model for b_{ext}

Whilst we do not usually observe $b_{\text{ext}}(h)$ it is still useful to specify an observation model for the extinction coefficient. Indeed, at inference stage, $\varphi(x|h)$ might fail to capture observational noise which will inevitably hinder variance calibration for the predicted extinction coefficient profile distribution. Using an observation model over b_{ext} introduces an additional degree of freedom that will help calibrate the variance of observations.

We propose a simple log-normal observation model with mean $\varphi(x|h)$ for the extinction coefficient,

$$b_{\text{ext}}(h)|\varphi(x|h) \sim \mathcal{LN}\left(\log \varphi(x|h) - \frac{\sigma_{\text{ext}}^2}{2}, \sigma_{\text{ext}}\right), \tag{6}$$

where $\sigma_{\text{ext}} > 0$ is a scale parameter. This observation model conserves the mean² from our prior $\varphi(x|h)$ and simply includes observational noise through the scale parameter σ_{ext} . When extinction coefficient observations are available, σ_{ext} can be calibrated following the procedure described in Section 3.3.3.

We emphasize that tuning σ_{ext} requires only a small number of extinction coefficient observations. This is in contrast to the AOD observations, which constitute the principal source of data we observe in the context of this work.

² $\mathbb{E}[b_{\text{ext}}(h)] = \mathbb{E}[\mathbb{E}[b_{\text{ext}}(h)|\varphi(x|h)]] = \mathbb{E}\left[\exp\left(\log \varphi(x|h) - \frac{\sigma_{\text{ext}}^2}{2} + \frac{\sigma_{\text{ext}}^2}{2}\right)\right] = \mathbb{E}[\varphi(x|h)]$.

2.2.3. An observation model for τ

AOD observations constitute the primary source of information on aerosols optical properties in the context of this work since it is collected routinely at a global scale by satellite products. Satellite retrievals are however prone to distortion, that may arise from observational noise (Remer et al., 2005) or assumptions made in the AOD retrieval algorithm (Mielonen et al., 2011; Wu et al., 2016). Therefore, as for the extinction coefficient, we propose to model these distortions using a log-normal observation model for the AOD conditional on $\varphi(x|h)$.

Formally, let $\eta = \int_0^H \varphi(x|h)dh$ be the column integration of our prior, then we propose a mean-parametrized observation model for the AOD conditional on η , given by

$$\tau|\eta \sim \mathcal{LN}\left(\log \eta - \frac{\sigma^2}{2}, \sigma\right), \tag{7}$$

$$\eta = \int_0^H \varphi(x|h)dh, \tag{8}$$

with a scale parameter $\sigma > 0$ which can be tuned against AOD observations jointly with other model parameters following the procedure described in Section 2.3.

When working with multiple AOD observations τ_1, \dots, τ_n , the scale parameter σ will be assumed shared among atmospheric columns. The mean parameter η (or location parameter μ for the canonical parametrization) will however be column-specific.

2.3. Tuning and posterior inference

2.3.1. Finite-sample problem formulation

We now make things concrete and assume we observe the AOD for n columns, which we stack into the vector $\tau = [\tau_1 \dots \tau_n]^T \in \mathbb{R}^n$. For the i th column, we also observe m_i vertically resolved meteorological covariates $x_i^{(1)}, \dots, x_i^{(m_i)}$ and their respective altitudes $h_i^{(1)} < \dots < h_i^{(m_i)}$, such that $x_i^{(j)} \sim p(x|h_i^{(j)})$.

We concatenate these observations into a dataset $\mathcal{D} = \left\{ \left(x_i^{(j)}, h_i^{(j)} \right)_{j=1}^{m_i}, \tau_i \right\}_{i=1}^n$ and denote $M = \sum_{i=1}^n m_i$ the total number of vertically resolved samples. The model description for the i th column is summarized in Figure 2.

Our objective is to use \mathcal{D} to learn the mapping $\varphi(x|h)$. Within the Bayesian framework, this objective is twofold:

- (A) Update the prior placed on φ with the observations from \mathcal{D} . This corresponds to computing the posterior distribution of φ given τ .
- (B) Tune the model hyperparameters by maximizing the marginal log-likelihood $\log p(\tau)$. The hyperparameters include the log-normal scale parameter σ and any parameter from the GP mean m and kernel k .

Regarding point (A), since φ results from a transformation of GP f , we will rather focus on the posterior distribution of the GP directly for convenience. If $f \in \mathbb{R}$ denotes a realization of f at any input $x|h$, the posterior distribution of $f(x|h)$ given observations is denoted $p(f|\tau)$. Having access to the posterior $p(f|\tau)$ allows computation of the predictive mean and variance of $\varphi(x|h)$ following

$$\mathbb{E}[\varphi(x|h)|\tau] = \int_{\mathbb{R}} \psi(f) e^{-h/L} p(f|\tau) df, \tag{9}$$

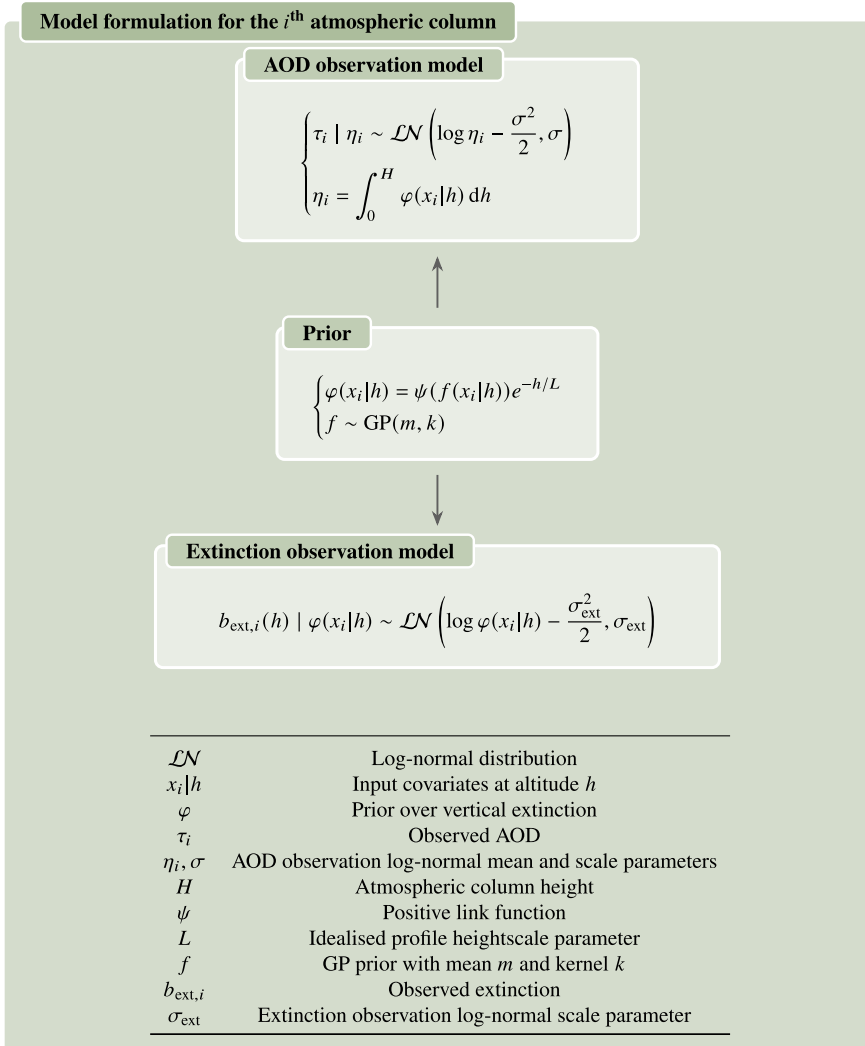


Figure 2. Observation models and prior formulation for the i^{th} atmospheric column. The model follows a hierarchical Bayesian structure represented by the arrows. We first specify the prior, and then express the observation models conditionally on the prior.

$$\text{Var}(\varphi(x|h)|\tau) = \mathbb{E}\left[\varphi(x|h)^2|\tau\right] - \mathbb{E}[\varphi(x|h)|\tau]^2. \tag{10}$$

The above can be estimated with Monte Carlo by drawing samples from the posterior $p(f|\tau)$. In Section 3.3.1, we will obtain a simpler closed-form solution for the particular choice $\psi = \exp$.

Regarding point (B), the marginal log-likelihood $\log p(\tau)$ is unfortunately intractable. Indeed, let $\mathbf{x} = [x_1^{(1)} \dots x_n^{(m_n)}]^\top \in \mathcal{X}^M$ denote the concatenation of all input entries from the dataset and let $\mathbf{f} \in \mathbb{R}^M$ denote a realization of $f(\mathbf{x})$. Assuming independence of AOD observations τ_1, \dots, τ_n conditionally on the GP realization over atmospheric columns \mathbf{f} , the marginal likelihood $p(\tau)$ is expressed in terms of the observation model and prior distributions following

$$p(\tau) = \int_{\mathbb{R}^M} p(\tau|\mathbf{f})p(\mathbf{f})d\mathbf{f} \quad \text{with} \quad p(\tau|\mathbf{f}) = \prod_{i=1}^n p(\tau_i|\mathbf{f}_i), \tag{11}$$

where $\mathbf{f}_i = [\mathbf{f}_i^{(1)} \dots \mathbf{f}_i^{(m_i)}]^\top \in \mathbb{R}^{m_i}$ corresponds to the GP realization over the i th column only. Because the observation model is log-normal, this integral is however intractable. To circumvent this intractability, we propose to tackle objective (B) by maximizing a proxy of the marginal log-likelihood presented in Section 2.3.3.

2.3.2. A sparse variational approximation of the posterior

We start by addressing objective (A). The predictive posterior distribution of interest is given by

$$p(\mathbf{f}|\tau) = \frac{p(\tau|\mathbf{f})p(\mathbf{f})}{\int_{\mathbb{R}} p(\tau|\mathbf{f})p(\mathbf{f})d\mathbf{f}}. \tag{12}$$

The integral denominator is not available in closed form, making the posterior intractable. We propose to use a variational approximation scheme (Titsias, 2009; Matthews et al., 2016; Leibfried et al., 2020) to substitute this intractable inference problem with a tractable optimization problem. In addition, we make the variational approximation sparse (Titsias, 2009) such that the model can scale to large amounts of data.

Let $\mathbf{z} = [z_1 \dots z_p]^\top \in \mathcal{X}^p$ be a set of $p \ll M$ inducing locations over the space of inputs. Their evaluation by the GP follows a multivariate normal distribution $f(\mathbf{z}) \sim \mathcal{N}(0, \mathbf{K}_{\mathbf{z}\mathbf{z}})$, where $\mathbf{K}_{\mathbf{z}\mathbf{z}} = k(\mathbf{z}, \mathbf{z})$. We denote $\mathbf{u} = f(\mathbf{z}) \in \mathbb{R}^p$ and refer to this vector as *inducing variables*.

A p -dimensional parametric distribution is set over these inducing variables. We choose this distribution as a multivariate normal defined by $q(\mathbf{u}) := \mathcal{N}(\mathbf{u}|\boldsymbol{\mu}_{\mathbf{z}}, \boldsymbol{\Sigma}_{\mathbf{z}})$. $\boldsymbol{\mu}_{\mathbf{z}} \in \mathbb{R}^p, \boldsymbol{\Sigma}_{\mathbf{z}} \in \mathbb{R}^{p \times p}$ are called the *variational parameters* and need to be tuned such that $q(\mathbf{u})$ best approximates the true posterior $p(\mathbf{u}|\tau)$.

Once this is achieved, we take as an approximation to $p(\mathbf{f}|\tau)$ the variational posterior defined by $q(\mathbf{f}) := \int_{\mathbb{R}^p} p(\mathbf{f}|\mathbf{u})p(\mathbf{u})d\mathbf{u}$, which is given in closed form by

$$q(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\bar{\boldsymbol{\mu}}_{x|h}, \bar{\boldsymbol{\Sigma}}_{x|h}), \tag{13}$$

$$\bar{\boldsymbol{\mu}}_{x|h} = k(x|h, \mathbf{w})\mathbf{K}_{\mathbf{z}\mathbf{z}}^{-1}\boldsymbol{\mu}_{\mathbf{z}}, \tag{14}$$

$$\bar{\boldsymbol{\Sigma}}_{x|h} = k(x|h, x|h) - k(x|h, \mathbf{z})(\mathbf{K}_{\mathbf{z}\mathbf{z}}^{-1} - \mathbf{K}_{\mathbf{z}\mathbf{z}}^{-1}\boldsymbol{\Sigma}_{\mathbf{z}}\mathbf{K}_{\mathbf{z}\mathbf{z}}^{-1})k(\mathbf{z}, x|h). \tag{15}$$

Naturally, (13) can be extended to describe a variational posterior over multiple GP entries. Namely, let $\mathbf{x}^* \in \mathcal{X}^D$ denote a vector of input entries. If $\mathbf{f}^* \in \mathbb{R}^D$ denotes a realization of $f(\mathbf{x}^*)$, then the associated variational posterior is given by

$$q(\mathbf{f}^*) = \mathcal{N}(\mathbf{f}^*|\bar{\boldsymbol{\mu}}_{\mathbf{x}^*}, \bar{\boldsymbol{\Sigma}}_{\mathbf{x}^*}), \tag{16}$$

$$\bar{\boldsymbol{\mu}}_{\mathbf{x}^*} = k(\mathbf{x}^*, \mathbf{w})\mathbf{K}_{\mathbf{z}\mathbf{z}}^{-1}\boldsymbol{\mu}_{\mathbf{z}} \in \mathbb{R}^D, \tag{17}$$

$$\bar{\boldsymbol{\Sigma}}_{\mathbf{x}^*} = k(\mathbf{x}^*, \mathbf{x}^*) - k(\mathbf{x}^*, \mathbf{z})(\mathbf{K}_{\mathbf{z}\mathbf{z}}^{-1} - \mathbf{K}_{\mathbf{z}\mathbf{z}}^{-1}\boldsymbol{\Sigma}_{\mathbf{z}}\mathbf{K}_{\mathbf{z}\mathbf{z}}^{-1})k(\mathbf{z}, \mathbf{x}^*) \in \mathbb{R}^{D \times D}. \tag{18}$$

The sparse nature of this approach becomes apparent in (17) and (18). Indeed, regardless of the number of samples we wish to evaluate the variational posterior over, we only need to invert a $p \times p$ matrix, incurring a $\mathcal{O}(p^3)$ computational cost.

2.3.3. Learning the variational parameters $\boldsymbol{\mu}_{\mathbf{z}}, \boldsymbol{\Sigma}_{\mathbf{z}}$

As mentioned above, the variational parameters $\boldsymbol{\mu}_{\mathbf{z}}, \boldsymbol{\Sigma}_{\mathbf{z}}$ need to be tuned such that $q(\mathbf{u})$ best approximates the posterior $p(\mathbf{u}|\tau)$, which is intractable. This problem is casted as the maximization of an objective called the evidence lower-bound (ELBO), given by

$$\text{ELBO}(q) = \mathbb{E}_{q(\mathbf{f})}[\log p(\tau|\mathbf{f})] - \text{KL}(q(\mathbf{u})\|p(\mathbf{u})). \tag{19}$$

The ELBO is a lower-bound to the marginal log-likelihood $\log p(\tau) \geq \text{ELBO}(q)$. It can thus be used as a proxy of $\log p(\tau)$ to also tune the model hyperparameters and fulfill objective (B).

The second term in (19) is the Kullback–Leibler divergence between two multivariate normal distributions. It admits a closed-form expression and can be computed. The first term, on the other hand, is an expected log-likelihood under the variational posterior which cannot be analytically computed. It can be decomposed into column-wise terms

$$\mathbb{E}_{q(\mathbf{f})}[\log p(\tau|\mathbf{f})] = \sum_{i=1}^n \mathbb{E}_{q(\mathbf{f}_i)} \left[\log \mathcal{LN} \left(\tau_i \mid \log \eta_i - \frac{\sigma^2}{2}, \sigma \right) \right]. \tag{20}$$

To estimate (20), we must first evaluate the mean parameter $\eta_i = \int_0^H \psi(f(x_i|h)) e^{-h/L} dh$ with the finite number of GP evaluations \mathbf{f}_i we have access to. We propose to use the trapezoidal integration scheme given by

$$\hat{\eta}_i = \sum_{j=1}^{m_i-1} \frac{\psi(\mathbf{f}_i^{(j+1)}) e^{-h_i^{(j+1)}/L} - \psi(\mathbf{f}_i^{(j)}) e^{-h_i^{(j)}/L}}{2} (h_i^{(j+1)} - h_i^{(j)}). \tag{21}$$

While we choose the trapezoidal rule for simplicity, we note that alternative finite-sample integration schemes can be chosen here in accordance with the needs.

Second, because of the log-normal observation model, the expected log-likelihood remains intractable. To estimate it, while allowing backpropagation through the variational parameters, we use a reparameterization trick (Kingma et al., 2015). Namely, we sample $\epsilon_i \sim \mathcal{N}(0, \mathbf{I}_{m_i})$ and compute $\mathbf{f}_i = \bar{\boldsymbol{\mu}}_i + \bar{\boldsymbol{\Sigma}}_i^{1/2} \epsilon_i$, where $\bar{\boldsymbol{\mu}}_i, \bar{\boldsymbol{\Sigma}}_i$ are the variational posterior parameters for the i th column and are obtained by application of (17) and (18) over the predictors of the i th column. The resulting GP sample \mathbf{f}_i is then used to estimate the mean parameter $\hat{\eta}_i$ following (21) and we can approximate the expected log-likelihood with its one-sample estimate

$$\mathbb{E}_{q(\mathbf{f}_i)} \left[\log \mathcal{LN} \left(\tau_i \mid \log \eta_i - \frac{\sigma^2}{2}, \sigma \right) \right] \approx \log \mathcal{LN} \left(\tau_i \mid \log \hat{\eta}_i - \frac{\sigma^2}{2}, \sigma \right). \tag{22}$$

This method allows estimation of the ELBO objective, which in turn can be maximized with respect to the variational parameters $\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z$ using a stochastic gradient approach for example. As mentioned above, the model hyperparameters can also be tuned jointly with this objective, hence fulfilling objective (B). These include the log-normal scale σ or the kernel k hyperparameters (e.g., variances and lengthscales), with an option to parametrize these kernels using feature maps given by deep neural networks (Law et al., 2019). As it is standard in sparse variational GPs (Titsias, 2009), we also learn the inducing locations \mathbf{z} .

3. Experimental setup

Our motivating application is to reconstruct aerosol vertical profiles from remote-sensing AOD observations. However, to validate the proposed methodology, we also need to observe extinction coefficient profiles. While the latter can be collected from vertically resolved remote sensing instruments (Holben et al., 1998; Winker et al., 2013), the collocation of measurements from different devices raises non-trivial questions regarding the self-consistency of the collocated dataset—and thus the validation procedure. In contrast, climate models are self-consistent and offer readily available simulations of gridded AOD and extinction coefficient profiles. We hence propose to evaluate our model using observations from a global climate model, the ECHAM-HAM global aerosol-climate model simulation (Stier et al., 2005, 2007; Zhang et al., 2012).

The objective of this experiment is to apply our model to the vertical disaggregation of the AOD simulated with ECHAM-HAM, and compare our prediction to the extinction coefficient simulated with ECHAM-HAM. In this section, we first describe the ECHAM-HAM dataset used in the experiments. We then outline the model setup and the inference procedure. Finally, we present the evaluation metrics used to quantify the quality of the reconstructed vertical profiles.

3.1. Dataset

3.1.1. The ECHAM-HAM dataset

The aerosol-climate model ECHAM-HAM is a self-consistent global climate model of aerosol radiative properties and CCN which demonstrates excellent agreement with AOD measurements from ground-based sun-photometers and satellite retrievals (Stier et al., 2005). It computes the evolution of log-normal aerosol mass and number modes—for species sulfates, black carbon, organic carbon, sea salt, and mineral dust—by taking into account physical and chemical particle processes. The simulation used includes aerosol optical properties, aerosol tracers, and meteorological variables at a $1.8^\circ \times 1.8^\circ$ horizontal resolution on a Gaussian grid and over 31 levels of vertical resolution and for 8 regularly spaced time steps over a day (06/06/2008). The resulting dataset counts 147,456 atmospheric columns used as training points. Table 1 provides its detailed dimensions.

The dataset captures two important aspects of aerosol properties variability: (i) because the dataset is global, it captures the spatial variability of aerosols, which is arguably the greatest contribution to the overall variability of aerosol properties. Indeed, the distribution of particles in the troposphere is highly heterogeneous due to the combination of spatially diverse sources and the relatively short lifetime of particles and gases (Carslaw and Pringle, 2022); (ii) the dataset covers a full day, which provides a good coverage of the diurnal cycle of aerosols.

However, a single day of data may fail to account for the seasonal variations of aerosol properties. The seasonal variability of aerosols is strongly regionally dependent, where variations in aerosol concentration can exceed a factor of 10 in regions prone to episodic emissions (e.g., wildfires, human activity), but vary only by 10% over a year in regions with persistent atmospheric circulation such as the Pacific subtropics (Carslaw and Pringle, 2022). Therefore, we expect that the demonstrated predictive performance of our model over this dataset should carry over to more stable regions for different seasons, but it remains uncertain from experiments on this dataset whether the model may generalize to the regional temporal variability induced by episodic events.

3.1.2. Choice of variables

ECHAM-HAM offers simulations of aerosol properties at particular wavelengths that match the wavelengths at which operate specific instruments for evaluation. We use the ECHAM-HAM AOD at 550 nm as the response variable to vertically disaggregate and the ECHAM-HAM extinction coefficient at 533 nm

Table 1. Gridded variables from ECHAM-HAM simulation data

	Name	Notation	Dimensions
Predictors	Temperature	T	$(t, \text{lat}, \text{lon}, \text{lev})$
	Pressure	P	$(t, \text{lat}, \text{lon}, \text{lev})$
	Relative humidity	RH	$(t, \text{lat}, \text{lon}, \text{lev})$
	Vertical velocity	ω	$(t, \text{lat}, \text{lon}, \text{lev})$
Response	AOD 550 nm	τ	$(t, \text{lat}, \text{lon})$
Groundtruth	Extinction coefficient 533 nm	b_{ext}	$(t, \text{lat}, \text{lon}, \text{lev})$

Note. The grid includes 8 time steps (t), 96 latitude levels (lat), 192 longitude levels (lon), and 31 vertical pressure levels (lev)—which is a proxy for h . Our objective is to vertically disaggregate the response τ using the vertically resolved predictors (T, P, RH, ω) and spatiotemporal columns locations ($t, \text{lat}, \text{lon}$).

as the groundtruth variable to evaluate our predictions against. The AOD and the extinction coefficients are simulated in ECHAM-HAM at slightly different wavelengths in order to match the wavelength at which operate the respective sensors (550 nm for the MODIS AOD product; Remer et al., 2005 and 532 nm for the CALIOP extinction; Winker et al., 2013).

In order to make the model easily applicable to any satellite AOD retrieval, we select vertically resolved predictors among standard meteorological variables, that could in practice be easily obtained from reanalysis data: $T, P, RH,$ and ω . By limiting ourselves to these variables, which are comprehensively covered by reanalysis products, we favor a global applicability of the model to AOD observations from most times and locations.

In what follows, the input variable writes $x = (t, \text{lat}, \text{lon}, T, P, RH, \omega)$. We standardize meteorological predictors and transform them using a rank-based inverse normal mapping to achieve approximately normal distributions.

3.2. Model setup and tuning

3.2.1. Choice of idealized profile height scale L

A simple choice for the idealized profile height scale is the boundary layer upper altitude, typically taken at 2 km. This is the value assumed by MISR, VIIRS, and MODIS C6_DT AOD retrieval algorithms (Levy et al., 2007; Kahn and Gaitley, 2015; Laszlo and Liu, 2016). In our experiment, we also choose to set $L = 2$ km.

3.2.2. Choice of mean and covariance structure

We set the GP with a zero mean function $m = 0$ and kernel taken as the sum of a spatiotemporal and a meteorological contribution to the covariance

$$k(x|h, x'|h') = \gamma_1 k_{\text{ST}}(\{t, \text{lat}, \text{lon}\}, \{t', \text{lat}', \text{lon}'\}) + \gamma_2 k_{\text{meteo}}(\{T, P, \omega, RH\}|h, \{T', P', \omega', RH'\}|h'), \quad (23)$$

with respective variances γ_1 and γ_2 . We propose to parametrize the kernels using the Matérn family, which has been widely used to work with spatiotemporal data (Stein, 1999). We denote C_ν the Matérn- ν covariance function with automatic relevance determination (that is independent lengthscales for each entry). This choice of Matérn order ν guarantees a desired level of regularity for the resulting GP. Details on the Matérn covariance functions are provided in Appendix A.

The spatiotemporal kernel is taken as the product kernel

$$k_{\text{ST}}(\{t, \text{lat}, \text{lon}\}, \{t', \text{lat}', \text{lon}'\}) = C_{3/2}(|t - t'|) C_{3/2} \left(\text{Hav} \left(\begin{bmatrix} \text{lat} \\ \text{lon} \end{bmatrix}, \begin{bmatrix} \text{lat}' \\ \text{lon}' \end{bmatrix} \right) \right), \quad (24)$$

where Hav denotes the Haversine distance—or great-circle distance—that determines the distance between two points on a sphere given their latitudes and longitudes. Each Matérn covariance admits a length scale, respectively denoted ℓ_t and $\ell_{\text{lat}, \text{lon}}$. Note that $C_{3/2}$ may not be positive definite over Hav for certain lengthscales (Feragen and Hauberg, 2016; Borovitskiy et al., 2020), thus we set an upper bound on $\ell_{\text{lat}, \text{lon}}$ to ensure positive-definiteness. The Matérn order $\nu = 3/2$ ensures the GP is continuously differentiable with respect to time and space.

The spatiotemporal kernel (24) makes the predicted extinction smooth across time and space. Its product structure ensures that for distant times (t, t') or distance locations ($(\text{lat}, \text{lon}), (\text{lat}', \text{lon}')$), the spatiotemporal covariance vanishes to zero. Hence, the GP covariance of predictions distant in space or time will disregard spatiotemporal predictors. This prevents overfitting over spatiotemporal predictors.

The meteorological kernel is taken as an automatic relevance determination covariance

$$k_{\text{meteo}}(\{T, P, \omega, RH\}|h, \{T', P', \omega', RH'\}|h') = C_{1/2} \left(\begin{bmatrix} |T - T'| \\ |P - P'| \\ |RH - RH'| \\ |\omega - \omega'| \end{bmatrix} \right), \quad (25)$$

where each dimension admits its own length scale parameter, respectively denoted $\ell_T, \ell_P, \ell_{RH}$, and ℓ_ω (we permit ourselves to drop the conditioning on h and h' for conciseness). The Matérn order $\nu = 1/2$ ensures that the GP is continuous with respect to meteorological predictors.

The kernel on meteorological predictors (25) allows to GP to learn how meteorological predictors should modulate extinction. It introduces covariance between predicted extinctions if meteorological predictors are close, even if the spatiotemporal covariance is zero (because of the additive structure in (23)). In practice, we find that setting a unit variance for this kernel $\gamma_2 = 1$ helps prevent vanishing kernel signal when tuning the hyperparameters.

We include in Appendix B a comparison of this choice of covariance structure with a more conventional tensor form covariance and evaluates their respective predictive performance.

3.2.3. Sparse variational GP setup and tuning

We use $p = 200$ inducing locations \mathbf{z} as an arbitrary choice (fewer or more inducing locations can be used following needs). They are initialized by randomly drawing samples at boundary layer altitude from \mathbf{x} . The variational distribution $q(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$ is initialized with $\boldsymbol{\mu}_z = \mathbf{0}$ and $\boldsymbol{\Sigma}_z = \mathbf{I}_p$. We denote $\Theta = \{\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z, \mathbf{z}, \sigma, \gamma_1, \gamma_2, \ell_t, \ell_{lat,lon}, \ell_T, \ell_P, \ell_{RH}, \ell_\omega\}$ the joint set of variational and model hyperparameters. We maximize the ELBO objective (19) with respect to Θ using the Adam optimizer (Kingma and Ba, 2015).

Since the dataset includes $8 \times 96 \times 192 = 147,456$ atmospheric columns, evaluating the ELBO for all columns at each optimization step is computationally prohibitive. Instead, we use a stochastic gradient approach and sample random mini-batches of columns over which we compute an unbiased noisy estimate of the ELBO gradient. Whilst compromising statistical efficiency, the high variance estimates of gradients have regularizing properties that mitigate overfitting, and this approach is guaranteed to converge to a local optima for Θ (Hoffman et al., 2013). To meet memory limitations, we set a mini-batch size of 64, and therefore choose a small learning rates of 0.01 to avoid taking large steps based on a few samples. When possible, a larger mini-batch size couple with a larger learning rate should allow to reach a better local optimum faster (Hoffman et al., 2013).

3.3. Inference procedure

3.3.1. Exponential link

We choose an exponential link function $\psi = \exp$ to warp the GP. This particular choice is mathematically convenient as it makes the prior $\varphi(x|h)$ a log-normal random variable given by

$$\varphi(x|h) = e^{f(x|h)-h/L} \sim \mathcal{LN}\left(m(x|h) - h/L, \sqrt{k(x|h, x|h)}\right). \tag{26}$$

A direct consequence is that we know the analytical expressions of its moments and quantiles, which alleviates the need for estimation procedures. This becomes of particular interest when making prediction, since the variational posterior that approximates $\varphi(x|h)|\tau$ is also a log-normal distribution given by $\mathcal{LN}\left(\bar{\mu}_{x|h} - h/L, \bar{\Sigma}_{x|h}^{1/2}\right)$ following notations from Section 2.3.2.

Therefore, we obtain approximations of the posterior mean, variance, and quantiles in closed-form following

$$\mathbb{E}[\varphi(x|h)|\tau] \approx \exp\left(\bar{\mu}_{x|h} - \frac{h}{L} + \frac{1}{2}\bar{\Sigma}_{x|h}\right), \tag{27}$$

$$\text{Var}(\varphi(x|h)|\tau) \approx \left(e^{\bar{\Sigma}_{x|h}} - 1\right) \exp\left(2\bar{\mu}_{x|h} - \frac{2h}{L} + \bar{\Sigma}_{x|h}\right), \tag{28}$$

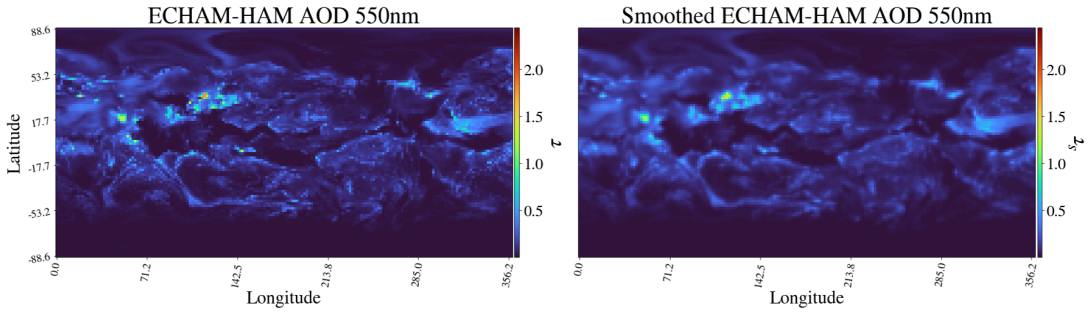


Figure 3. Left: ECHAM-HAM 550 nm AOD. Right: Spatially smoothed ECHAM-HAM 550 nm AOD.

$$Q_\alpha(\varphi(x|h)|\tau) \approx \exp\left(\bar{\mu}_{x|h} - \frac{h}{L} + \sqrt{2\bar{\Sigma}_{x|h}} \operatorname{erf}^{-1}(2\alpha - 1)\right), \tag{29}$$

where Q_α denotes the α -quantile function and erf denotes the error function. This allows for an efficient computation of the predicted mean extinction profile and confidence regions.

3.3.2. Rescaling the predicted profiles

The posterior distribution $p(\mathbf{f}|\tau)$ updates the GP with the information that the mapping φ should (in expectation) vertically integrate to the observed AOD τ . While instructive, this is a weak constraint on the prior and does not guarantee that the prediction will effectively integrate to the AOD of the atmospheric column.

To get stronger enforcement of column-integrated values and ensure predictions effectively integrate to the AOD in expectation, we propose to rescale the model against observed AOD. Because observations τ_1, \dots, τ_n are corrupted with high-frequency observational noise, we introduce a spatially smoothed version of the AOD observations that filters out this noise. By applying a spatial Gaussian smoothing filter to the ECHAM-HAM AOD, we obtain spatially smoothed versions of the observations which we denote ${}^s\tau_1, \dots, {}^s\tau_n$. A comparison of the fields is displayed in Figure 3.

At inference time, we then substitute the predicted posterior of extinction with its rescaled version given by

$${}^s\varphi(x_i|h)|\tau = \frac{{}^s\tau_i}{\int_0^H \mathbb{E}[\varphi(x_i|h)|\tau] dh} \varphi(x_i|h)|\tau, \tag{30}$$

for the i th column. This ensures that, in expectation, ${}^s\varphi(x_i|h)|\tau$ effectively integrates along height to ${}^s\tau$. When choosing $\psi = \exp$, performing this rescale is equivalent to shifting the location of the posterior in (26) by $\log {}^s\tau_i - \log \int_0^H \mathbb{E}[\varphi(x_i|h)|\tau] dh$. The integral is approximated with a trapezoidal scheme.

3.3.3. From ${}^s\varphi(x|h)|\tau$ to $b_{\text{ext}}(h)|\tau$

The final stage of inference is to estimate the posterior distribution of b_{ext} following the observation model (6). Because the chosen observation model conserves the mean, the posterior mean of $b_{\text{ext}}(h)$ is the same as the posterior mean of ${}^s\varphi(x|h)$. Higher-order moments however cannot be obtained analytically and will require estimation by sampling from $b_{\text{ext}}(h)|\tau$. This can be achieved by sampling from ${}^s\varphi(x|h)|\tau$ and then plugging the samples in the observation model (6) to sample from $b_{\text{ext}}(h)|{}^s\varphi(x|h)$.

The observation model scale parameter σ_{ext} needs to be calibrated against observational noise. We set σ_{ext} to the value that minimizes the integrated calibration index (ICI) given by

$$\text{ICI} = \int_0^1 |N_{\sigma_{\text{ext}}}(\alpha) - (1 - \alpha)| d\alpha, \tag{31}$$

where $N_{\sigma_{\text{ext}}}(\alpha)$ is the percentage of ECHAM-HAM extinction coefficient observations that fall within the $(1 - \alpha)$ credible interval of the distribution of $b_{\text{ext}}|\tau$. The ICI is evaluated against $n_{\text{calib}} = 200$ random atmospheric columns from the dataset. These columns form a separate calibration set which is solely used to calibrate σ_{ext} . This separate calibration set is negligible in size in comparison with the 147,456 atmospheric columns available in the simulation.

3.4. Evaluation procedure

We compare the predicted extinction coefficient profile $b_{\text{ext}}|\tau$ to the extinction coefficient from ECHAM-HAM simulations. For clarity, we refer to our proposed methodology as *VAExtGP* (for Variational Aerosol Extinction GP) in the presentation of experimental results.

3.4.1. Baseline models

We use as a comparative baseline an idealized exponential profile, similar to the vertical profiles postulated in AOD retrieval algorithms. For a fair comparison, we rescale the baseline profiles such that their integrated values match the AOD, following the procedure described in Section 3.3.2. We complement this baseline with a log-normal observation model for the extinction coefficient and calibrate its scale parameter against a separate calibration following the procedure outlined in Section 3.3.3. This enables probabilistic evaluation of the extinction coefficient profiles predicted with the idealized baseline.

In addition, to examine the contribution to predictions of our modeling choices, we consider the following ablated versions of *VAExtGP*:

- *GP only*: we remove the idealized exponential vertical contribution to the prior. The prior is specified as the exponential of a GP only, that is, $\varphi(x|h) = \exp(f(x|h))$ with $f \sim \text{GP}(m, k)$.
- *ST only*: we only use spatiotemporal covariates as input variables for the model, that is, $x = (t, \text{lat}, \text{lon})$. The kernel is set to $k = k_{\text{ST}}$.
- *Meteo only*: we only use meteorological covariates as input variables for the model, that is, $x = (T, P, \text{RH}, \omega)$. The kernel is set to $k = k_{\text{meteo}}$.

3.4.2. Evaluation metrics

We use two categories of metrics to quantify the quality of the predicted vertical profiles: deterministic metrics, which compare only the posterior mean prediction to the groundtruth ECHAM-HAM extinction coefficient profiles, and probabilistic metrics which evaluate the entire posterior probability distribution against the extinction coefficients from ECHAM-HAM simulation and thus also assess the quality of the uncertainty quantification in the posteriors. The metrics used are detailed in Table 2.

Table 2. Evaluation metrics

	Metric	Description	Best when
Deterministic	RMSE	Root mean square error	Close to 0
	MAE	Mean absolute error	Close to 0
	Corr	Pearson correlation	Close to 1
	Bias	Mean bias	Close to 0
	Bias98	Bias in the 98th percentile	Close to 0
Probabilistic	ELBO	Evidence lower-bound of groundtruth b_{ext}	Higher
	Calib95	95% calibration score, that is, $N_{\sigma_{\text{ext}}}(\alpha = 0.05)$	Close to 95%
	ICI	Integrated calibration index—see (31)	Close to 0

Note. Deterministic metrics compare the predicted posterior mean $\mathbb{E}[b_{\text{ext}}|\tau]$ to the ECHAM-HAM extinction coefficient; Probabilistic metrics evaluate the complete predicted posterior probability distribution of $b_{\text{ext}}|\tau$ against ECHAM-HAM extinction coefficient.

For each metric, we compute scores for pixels along the entire atmospheric columns and for pixels lying within the boundary layer only (<2 km). Scores are averaged across all pixels.

4. Results

4.1. Comparison to an idealized exponential baseline

Table 3 shows that the posterior mean profile predicted with the proposed method outperforms the idealized exponential baseline across deterministic metrics. We note that evaluating over the entire column consistently yields better scores compared to evaluating only over the boundary layer. This is to be expected as the extinction coefficient outside the boundary layer tends to vanish to zero and most of the variability occurs within the boundary layer.

The improvement in RMSE and MAE against the idealized baseline is slightly more significant when computed within the boundary layer rather than over the entire column. This supports the idea that *VAExtGP* improves upon an idealized exponential profile at predicting extinction since the boundary layer concentrates most extinction variability. The improvement is greatest in Bias and Bias98, where the proposed model consistently improves over the idealized baseline by one or multiple orders of magnitude. In particular, the improvement in Bias98 suggests that *VAExtGP* better captures extreme extinction values.

When considering the evaluation over the entire column, the ELBO and ICI scores are comparable for both methods. However, within the boundary layer, *VAExtGP* outperforms the baseline, by a substantial margin in terms of ICI. This indicates that, within the boundary layer, the predicted posterior probability distribution of *VAExtGP* better captures the ECHAM-HAM extinction variability. The 95% calibration scores are similar for both methods, with a slight advantage for the idealized baseline.

4.2. Reconstructed vertical profiles

Figure 4 displays slices for a fixed latitude of the vertically resolved predictors used, the ECHAM-HAM extinction coefficient, and the predicted extinction coefficient with *VAExtGP*. For comparison, the prediction at the same latitude with the idealized exponential baseline and additional predictions with *VAExtGP* for different latitudes are provided in Appendix C.

We observe that our predicted mean profile is able to reconstruct extinction patterns that are visually very similar to the groundtruth extinction coefficient. In comparison with the profiled predicted with the idealized exponential baseline in Figure C1, the extinction profiles predicted with our method look much more realistic. This is encouraging considering that the only aerosol optical property used is the AOD. We also observe that extreme extinction coefficient values appear to be well captured within the 95% confidence region of the predicted posterior distribution.

Since aerosol water uptake is related to relative humidity, we observe a good capacity to recover patterns corresponding to extinction due to aerosol swelling at low altitudes in the boundary layer. This

Table 3. Scores of the idealized exponential baseline and *VAExtGP* for the task of predicting ECHAM-HAM extinction profiles

Region	Method	RMSE (10^{-5})	MAE (10^{-6})	Corr	Bias (10^{-6})	Bias98 (10^{-5})	ELBO ^a	Calib95 (%)	ICI (10^{-2})
Entire column	<i>Idealized</i>	4.10	6.65	0.51	-2.40	-4.08	13.1	96.0	5.05
	<i>VAExtGP</i>	3.21 ± 0.04	5.29 ± 0.06	0.73 ± 0.01	-0.005 ± 0.161	-0.482 ± 0.212	13.0 ± 0.1	93.5 ± 0.1	6.60 ± 0.45
Boundary layer	<i>Idealized</i>	7.55	16.8	0.54	-12.9	-11.7	10.2	93.5	19.1
	<i>VAExtGP</i>	5.91 ± 0.07	15.3 ± 0.1	0.72 ± 0.06	-0.404 ± 0.71	-4.53 ± 0.44	10.7 ± 0.1	98.7 ± 0.1	6.26 ± 0.34

Note. "Entire column" means scores are computed for every altitude level; "Boundary layer" means scores are computed for altitude levels of the boundary level only (<2 km); the best scores for each metric and region are highlighted in bold; we report 1 standard deviation.

^aFor *Idealized*, we report the marginal log-likelihood of ECHAM-HAM extinction under the predicted profiles instead of the ELBO. This is an upper bound to the ELBO, which could possibly be lower.

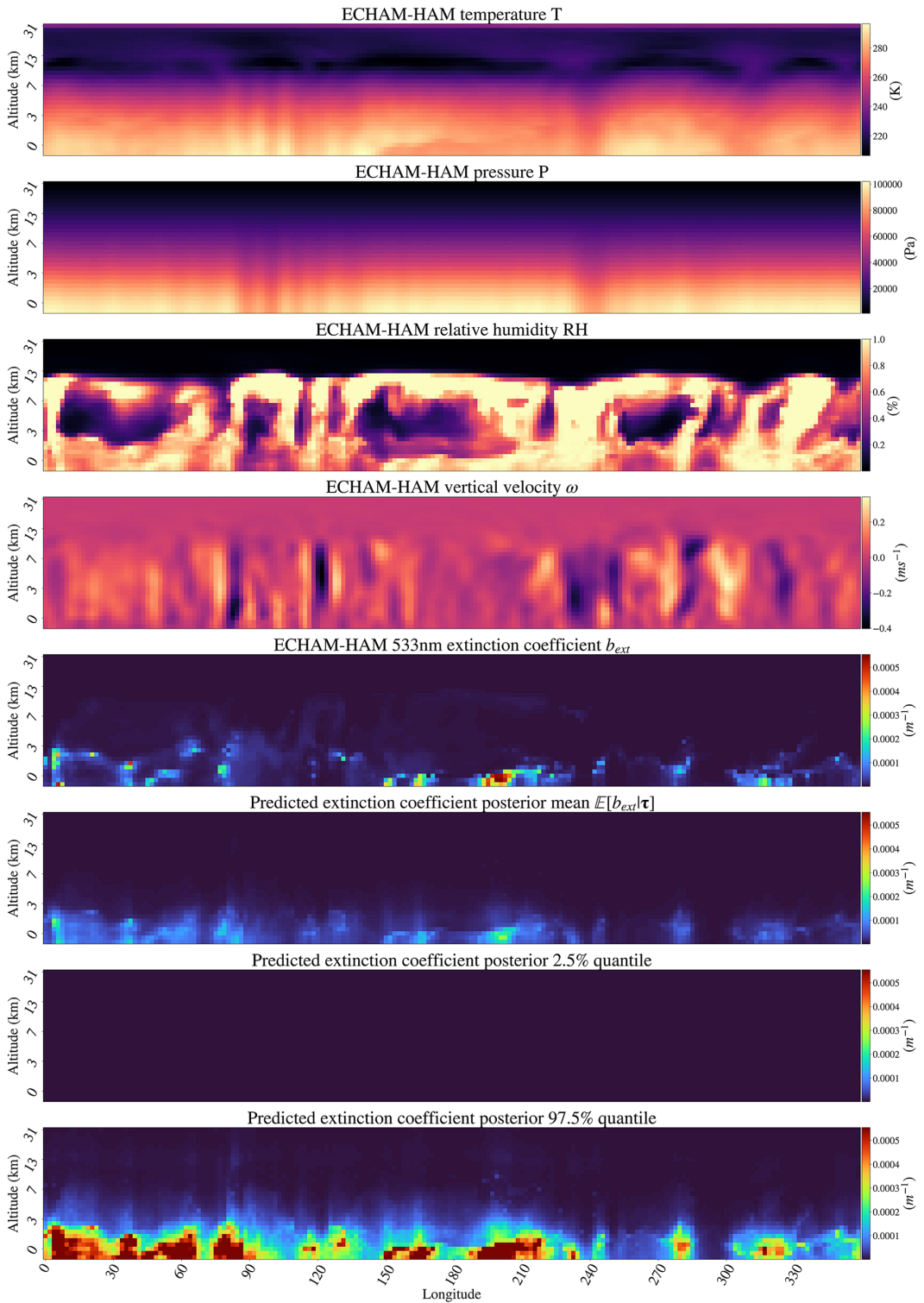


Figure 4. Vertical slices at latitude 51.29° of meteorological predictors (T, P, RH, ω), groundtruth extinction coefficient, predicted extinction coefficient posterior mean, 2.5% and 97.5% quantiles of the predicted extinction coefficient posterior distribution.

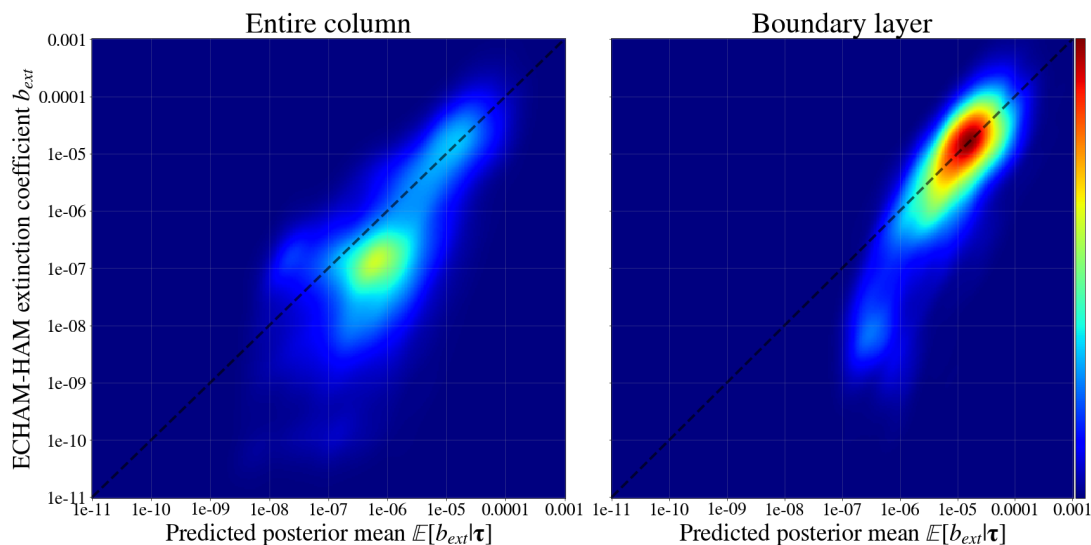


Figure 5. Density plots of groundtruth extinction coefficient values against predicted posterior mean extinction coefficient; Left: plotted for the entire column; Right: plotted for the boundary layer only; density plots are computed on a random subset on a random subset of 1000 samples drawn for the entire column (left) and in the boundary layer (right).

is particularly visible over the Pacific (longitudes 144° – 228°) and the Atlantic (longitudes 305° – 350°), where the predicted patterns of extinction tend to best reproduce the ECHAM-HAM extinction patterns.

However, the predicted mean profile faces greater difficulties to accurately reproduce the extinction patterns for longitudes 0° – 100° . This is particularly noticeable in Figure 4 around longitude 71.4° nearby Astana in Kazakhstan, and around longitude 17.3° in the Silesia region in Poland. The geographical locations suggest these are extinction patterns induced by human activity, which cannot be captured solely by meteorological variables. We conjecture these patterns are imputable to aerosols mass concentration, particles size, and radiative properties. These properties, unlike extinction due to swelling, are more complex and cannot be fully characterized by relative humidity, temperature, pressure, and updraft. It is worth highlighting that, whilst difficult to predict, these extinctions pattern appear nonetheless to be well captured within the 95% confidence region of the predicted posterior distribution.

In general, the posterior mean predictions tend to be smoother than the groundtruth extinction coefficient. The smoothing effect is a regularizing property of the GP prior, resulting in an overestimation of the extinction coefficient in certain regions with low extinction. For example, around longitude 200° in the Pacific, the ECHAM-HAM extinction in Figure 4 displays an extinction pocket with a sharp limit around altitude 1 km, above which extinction is virtually absent. In contrast, our predicted posterior mean is more diffuse and tends to spread out lightly above 1 km.

The density plots in Figure 5 support this observation, as a large mass of extinction coefficient around 10^{-7} m^{-1} tends to be overestimated by an order of magnitude. However, this trend is significantly reduced when focusing on the boundary layer, where most of the mass lies around the axis $y = x$. This suggests that most of the overestimation occurs above the boundary layer for low extinction ($< 10^{-6} \text{ m}^{-1}$). Within the boundary layer, however, the mean predictions are reasonably aligned with the groundtruth for high extinction coefficient values ($> 10^{-5} \text{ m}^{-1}$). Although low extinction coefficient values are less dominant in the boundary layer, they also tend to be overestimated. We attribute this to the smoothing effect of the GP prior, which is prone to overestimation between extinction pockets.

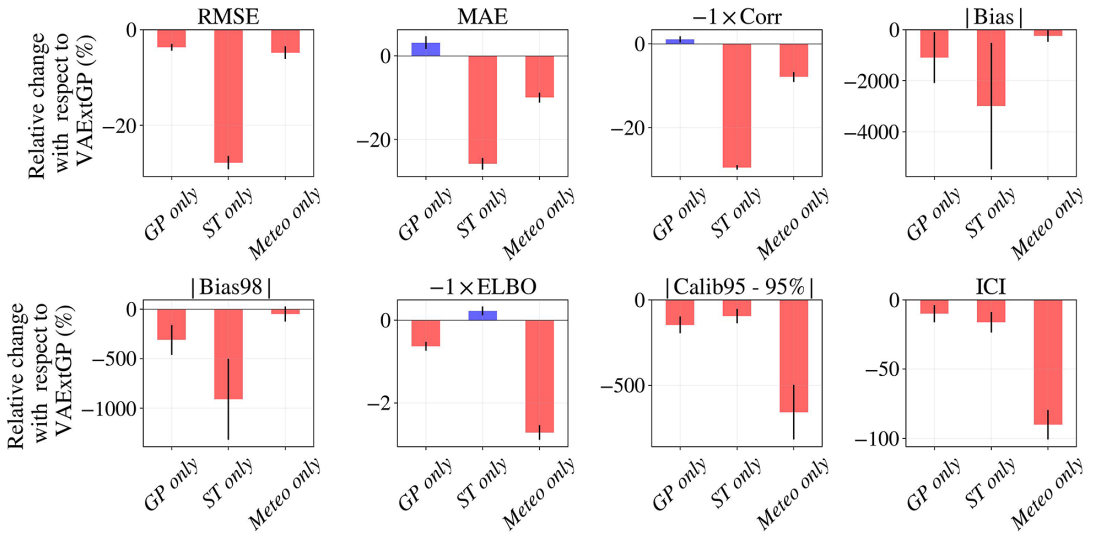


Figure 6. Relative changes in scores with respect to VAExtGP for the 3 ablated models: “GP only” (no idealized exponential term $e^{-h/L}$ in the prior), “ST only” (only spatiotemporal covariates in the input variable) and “Meteo only” (only meteorological covariates in the input variable). Red/Blue indicates that the performance is on average worse/better for the ablated model. We report 1 standard deviation.

4.3. Ablation study

We examine the contribution to predictions of our modeling choices by comparing the scores obtained with VAExtGP to the scores obtained with the three ablated models described in Section 3.4.1. Figure 6 shows the relative changes in scores with the ablated models, defined as

$$\% \text{Relative change} = \frac{\text{Score VAExtGP} - \text{Score ablated model}}{\text{Score VAExtGP}} \times 100, \tag{32}$$

where a positive value (blue) corresponds to an improvement for the ablated model, and a negative value (red) corresponds to a decrease in performance for the ablated model. Transformations of Corr, Bias, Bias98, ELBO, and Calib95 are taken such that smaller values correspond to better performance, and are detailed in Appendix C.

Overall, we find that excluding a component from the model leads to a decrease in performance across most metrics. Specifically, when using only a GP without the idealized exponential term in the prior (GP only), only spatiotemporal covariates (ST only), or only meteorological covariates (Meteo only), the model’s performance is adversely affected.

For deterministic metrics, the ablation of meteorological covariates (ST only) has a particularly pronounced impact, with drastic drop in Bias and Bias98. This highlights the key role of learning the mapping from meteorological variables to extinction profiles. For probabilistic metrics, we observe that the ablation of spatiotemporal covariates (Meteo only) leads to the largest decrease in scores, indicating that incorporating spatiotemporal information plays a significant role in calibrating the predicted posterior distribution. Furthermore, we note that whilst the performance of the GP only model is generally inferior to that of VAExt, emphasizing the importance of encoding prior knowledge about aerosol vertical distribution through the idealized exponential profile, it is also the ablated model that least degrades performance. This highlights the remarkable adaptivity of GPs, and supports their use to learn non-trivial relationship in data.

In conclusion, these results demonstrate the significance of incorporating both spatiotemporal and meteorological covariates, as well as the idealized exponential profile, in our proposed model.

4.4. Feature importance analysis

In this section, we investigate the role predictors in *VAExtGP* play in the prediction of the extinction profiles. We begin by examining the model hyperparameters to gain an understanding of how do predictors overall contribute to the covariance structure. We then ask the following questions:

- How do meteorological predictors modulate the predicted extinction?
- Do temperature and pressure provide valuable meteorological information for prediction or are they just proxies for height?

To address these questions, we propose a more detailed feature importance analysis that elucidates the contribution of predictors to individual predictions: Shapley values (SVs) (Shapley, 1953). Originally introduced as a game-theoretic concept, SVs have been widely adopted in machine learning to design local feature importance explanation models (Štrumbelj and Kononenko, 2014; Lundberg and Lee, 2017; Lundberg et al., 2018; Ghorbani and Zou, 2019). With SVs, we can estimate for each individual prediction, how much each predictor contributed, and hence explain predictions. In our study, we use the KernelSHAP SV-based explanation model (Lundberg and Lee, 2017).

4.4.1. Global analysis with model hyperparameters

Table 4 reports the hyperparameters values after tuning them to maximize the ELBO objective as described in Section 2.3.3. These values provide a global description of the importance of predictors contribution to the covariance structure. γ_1 and γ_2 respectively correspond to the variances of the spatiotemporal and meteorological kernels from (23). A larger variance hyperparameter indicates a greater contribution to the covariance. The lengthscales $\ell_{\text{lat,lon}}$, ℓ_T , ℓ_P , ℓ_{RH} , and ℓ_ω correspond to the lengthscales of each standardized predictor in the spatiotemporal kernel (24) and (25). A smaller length scale parameter means that the covariance is more sensitive to small variations in the predictors.

The value of γ_1 indicates that meteorological predictors globally contribute more to the covariance than spatiotemporal predictors. The large value of the temporal length scale ℓ_t (greater than 5) suggests a minimal sensitivity of the model to time. This is certainly because the dataset covers a single day and therefore exhibits limited temporal variability. In contrast, the small value of the spatial length scale $\ell_{\text{lat,lon}}$ indicates a high sensitivity to spatial location and suggests that spatial correlation occurs within a restricted spatial range. This likely arises from the spatial heterogeneity of aerosol concentration, resulting in localized, short-range spatial patterns.

Regarding meteorological predictors, the length scales indicate that—in general—predictions are more sensitive to variations in temperatures and less sensitive to variations in updraft. In fact, the large value of the updraft length scale ℓ_ω indicates, in general, a minimal correlation decay over the range of inputs.

4.4.2. How do meteorological predictors modulate the predicted extinction?

To gain a refined understanding of how meteorological predictors affect predictions, we conduct a SV analysis of their contribution to individual predictions. Since the meteorological predictors only modulate

Table 4. Hyperparameter values for *VAExtGP* tuned for standardized input variables following the procedure described in Section 2.3.3

γ_1	γ_2	ℓ_t	$\ell_{\text{lat,lon}}$	ℓ_T	ℓ_P	ℓ_{RH}	ℓ_ω
0.26 ± 0.02	1.0^\dagger	7.6 ± 0.2	0.85 ± 0.06	2.5 ± 0.1	3.4 ± 0.1	3.4 ± 0.1	6.6 ± 0.1

Note. Length scale is unitless and apply to standardized predictors; hyperparameters are tuned using 10 different initialization seeds; we report 1 standard deviation; \dagger indicates the value is fixed in the model and not tuned against data.

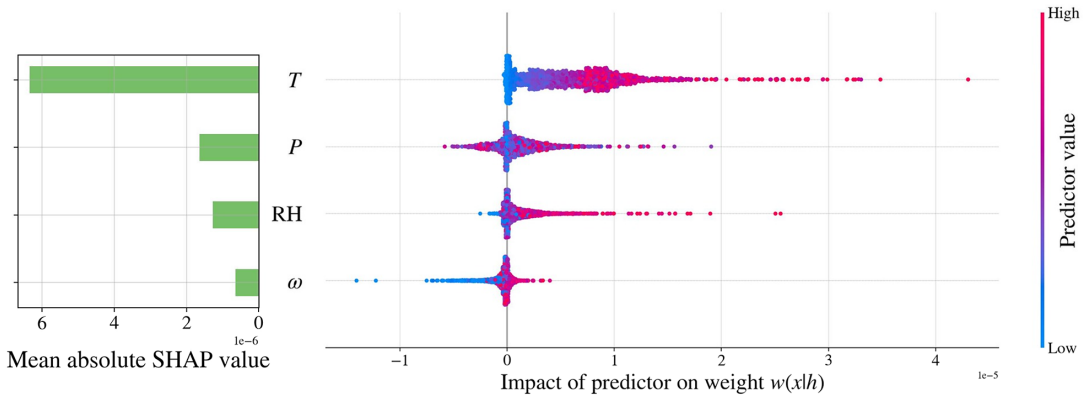


Figure 7. Left: Mean absolute Shapley values of meteorological predictors in the boundary layer; Right: Beeswarm plot of Shapley values of individual predictions in the boundary layer for each meteorological predictor; red/blue dots indicate a sample where the predictor value lies on the upper/lower end of its distribution; Shapley values are computed on a random subset of 2000 samples in the boundary layer.

predictions through the weighting function, we only study the impact meteorological predictors have on the weight $w(x|h)$ predicted for the posterior mean response. We also focus on samples within the boundary layer since this is where most of the variability happens. Figure 7 displays the mean absolute SV obtained for each meteorological predictor and the contribution each predictor has on a randomly selected set of individual predictions.

We observe that in absolute mean, temperature is the factor that most influences predictions of the weighting function in the boundary layer. This is sensible as temperature is a smooth spatial field that gradually decreases with altitude, making T characteristic of the altitude level. Since the altitude typically correlates with extinction values, T becomes an informative proxy of extinction. This is corroborated by the plot of contributions to individual predictions: greater temperature (red)—lower altitudes—contribute to a greater weight $w(x|h)$ and hence greater extinction prediction while lower temperature (blue)—higher altitudes—yield lower impact on the weight and hence lower extinction prediction. It is interesting to notice that while pressure is also a proxy of altitude just like temperature, its mean absolute contribution is lower and the individual contribution plot does not display a clear identifiable trend. This suggests that the information conveyed by pressure might be redundant with the information conveyed by temperature, and could possibly be discarded from the model without loss of predictive performance.

Whilst the mean absolute contribution of relative humidity seems marginal compared to temperature, we notice in individual contributions that when RH contributes to a greater weight $w(x|h)$ (points lying on the right), it consistently corresponds to locations with high RH. This supports the observation that our model captures extinction due to aerosol water uptake. Similarly, the mean absolute contribution of ω is relatively small, but a clear trend appears in the individual contributions: a positive updraft consistently corresponds to a positive contribution of ω to the weight function, whereas a negative updraft corresponds to a negative contribution to the weight function. This is consistent with intuition as strong updrafts lead to the humidification of aerosols, and an increase in water content in aerosols increases in turn the extinction due to aerosol water uptake. Conversely, a negative updraft should lead to lower humidification of aerosols, and therefore a decrease in extinction due to aerosols water uptake.

4.4.3. Temperature and pressure: proxies for height?

Temperature and pressure carry significant information about height, and may therefore inform *VAExtGP* primarily as a proxy for height. Indeed, the vertical slices of T and P in Figure 4 exhibit an approximately

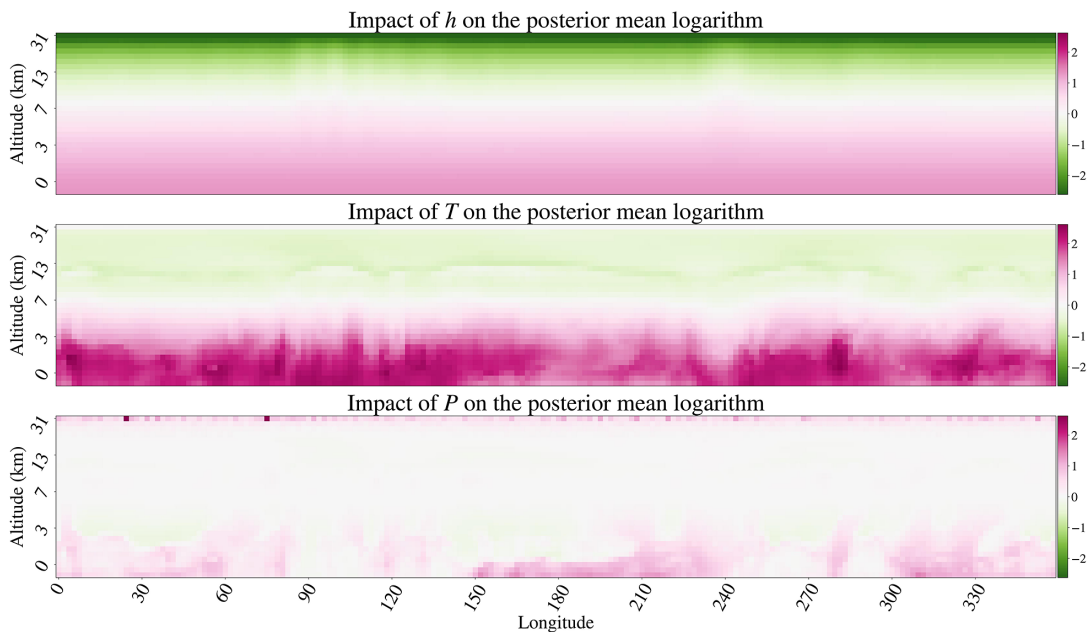


Figure 8. Vertical slices at latitude 51.29° (same as Figure 4) of the Shapley values for height h , temperature T , and pressure P ; the Shapley values are computed for predicting the logarithm of the posterior mean, that is, $\log \mathbb{E}[\varphi(x|h)|\tau]$; green/pink indicates a sample where the predictor has contributed to a lower/higher predicted extinction.

linear behavior with height. However, for a fixed height, temperature and pressure also exhibit variations along longitude, and may therefore provide additional information about meteorological conditions, beyond serving as a proxy for height. To investigate this, we conduct a SV analysis of the contribution of height, temperature, and pressure to the predicted logarithm of the posterior mean $\log \mathbb{E}[\varphi(x|h)|\tau]$. Figure 8 shows the reported SVs for the same vertical slice as Figure 4.

The contribution of h to predicted extinction is clearly visible, where greater altitude contributes to a lower predicted extinction and lower altitude contributes to a greater extinction. This corresponds to the domain knowledge we encode within the prior by using an idealized exponential component. The contribution of T displays a similar mode, in agreement with the analysis from Section 4.4.2: at high altitude, lower temperatures contribute to lower predicted extinction, whereas at low altitudes, higher temperatures contribute to greater predicted extinction. The contribution of P is in general marginal in comparison, which aligns with the conjecture that P carries redundant information and may be a confounding feature which can be discarded.

In the stratosphere (altitude > 10 km), the contribution of T becomes marginal in comparison with the contribution of h . This suggests that the contribution of the exponential profile dominates at high altitudes. In the troposphere (altitude < 10 km), the contribution of T dominates, indicating that temperature plays a key role in predictions at lower altitudes. This is particularly visible in the boundary layer (altitude $< 2 - 3$ km). Importantly, the temperature SVs display spatial heterogeneity within the boundary layer, where the contribution at a fixed altitude varies across longitude. This is different from the height SVs, and suggests that T informs the model with valuable meteorological information, beyond being a proxy for height.

To evaluate the contribution of T and P to prediction performance, we conduct ablation experiments where we train and evaluate *VAExtGP* by (i) removing P and (ii) removing both T, P from the input meteorological predictors. Table 5 shows the percentages of relative change in score in comparison with

Table 5. Percentages of relative change in scores in with respect to VAExtGP when removing P and (T,P) from the input meteorological predictors

Remove	RMSE	MAE	$-1 \times \text{Corr}$	Bias	Bias98	$-1 \times \text{ELBO}$	Calib95–95%	ICI
P	+2.7 ± 1.1	-0.77 ± 1.4	+1.5 ± 0.93	-52 ± 1.8 · 10 ²	+64 ± 27	+0.39 ± 0.13	+46 ± 18	+18 ± 7.7
T,P	-25 ± 2	-22 ± 1.6	-23 ± 0.02	-2.7 · 10 ³ ± 2.4 · 10 ³	-8.9 · 10 ² ± 4.2 · 10 ²	+8.0 · 10 ⁻² ± 1.6 · 10 ⁻¹	-23 ± 42	-7.9 ± 1.4 · 10 ¹

Note. Red/Blue indicates that the performance is worse/better for the ablated model with statistical significance at $p < 0.05$ level. We report 1 standard deviation.

the entire VAExtGP model. We follow the same conventions than in Section 4.3 and take transformations of Corr, Bias, Bias98, ELBO, and Calib95 such that smaller values correspond to better performance.

We find that: (i) predictive performance is overall improved by removing P . This supports the conjecture that P is a confounding factor, which carries redundant information for the prediction of extinction, and thereby is a source of noise in predictions. (ii) Removing both T and P overall harms predictive performance. This supports the observation that, whilst strongly correlated with height, temperature and pressure provide additional predictive utility.

5. Conclusion and outlooks

In this work, we introduce a GP-based methodology to vertically disaggregate the AOD using simple vertically resolved meteorological predictors such as temperature, pressure, or relative humidity. Our approach emphasizes uncertainty quantification using a Bayesian formalism. A successful application of our methodology to the vertical disaggregation of ECHAM-HAM simulated AOD is demonstrated. Our model outperforms an idealized baseline and displays capacity to recover realistic extinction patterns, in particular for extinction patterns arising from aerosol swelling in the boundary layer.

The simplicity of the explicit modeling assumptions grants better control and interpretability over the model and makes the choice of analysis strategy less subjective. While such simplicity can never account for the complex phenomena underpinning ACIs, this is balanced by a particular emphasis on the quantification of epistemic uncertainty through a principled Bayesian formalism.

Naturally, the modeling assumptions can be adapted to reflect different modeling choices. For example, where more relevant vertically resolved quantities can be obtained, they can seamlessly be incorporated into the predictors. The idealized exponential component of the prior can be replaced by an idealized Gaussian profile to reflect the assumptions made by different AOD retrieval algorithms (Wu et al., 2017; Li et al., 2020). The kernel design is also flexible and allows the user to specify a covariance structure that best accounts for pairwise dependencies of the predictors. Similarly, the positive link function, integration scheme, or number of inducing locations can be modified to fit specific needs.

Experiments demonstrate the proposed model is able to realistically reproduce vertical structures of extinction. In particular—for the chosen set of meteorological predictors—we are able to reliably predict extinction due to water vapor in the boundary layer. Because of the smoothness bestowed by the GP prior, the model tends to overestimate low extinction above the boundary layer and between extinction packets in the boundary layer. It is conjectured that the remaining unexplained extinctions patterns can be attributed to aerosols mass concentration, particles size, and radiative properties, which are more challenging to model and would require additional vertically resolved covariates. This constitutes an important area for future work.

Regarding methodology, several extensions remain to be explored. For example, while we only consider aerosol extinction at a single wavelength, the AOD is in general retrieved for multiple wavelengths ranging from 470 nm to 870 nm (Remer et al., 2005). By making the GP a function of the

wavelength $f(\lambda, x|h)$, it should be possible to leverage multiple AOD observations by introducing a notion of functional smoothness across the electromagnetic spectrum. Another exciting direction would be to allow the use of AOD observations which are not matched with vertically resolved predictors. This can be achieved by adapting the work of Chau et al. (2021) to our model, using a globally observed 2D field that would mediate the learning between unmatched AOD and vertically resolved predictors. Such addition would be particularly welcome as it would allow prediction of aerosol vertical profiles even at locations where AOD is not observed but vertically resolved predictors are.

The proposed method can also be useful to diagnose the extinction profiles produced by Earth system models (ESMs). For example, by training and validating our model on ESMs, one can produce a predictive model that learns ESM extinction characteristics. Predictions from this model can then be evaluated against real extinction observations, and thereby provide a quantitative indicator of whether ESM extinction characteristics match the characteristics of observed extinction.

Finally, a different exciting use case of the proposed methodology can be considered. If working with climate model data only, then one can choose to use any vertically resolved aerosol tracers as predictors. These include mass and number concentrations which are simulated for several aerosol modes (nucleation, aiten, coarse, and accumulation) and species (dust, sea salt, sulfate, black, and organic carbon). Leveraging the interpretability of the GP covariance function, the kernel hyperparameters can then indicate which mode/specie did contribute to predictions, hence providing a tool to better understand factors that modulate the optical properties of aerosols.

In future work, we intend to apply our model to AOD arising from satellite observations and meteorological predictors from reanalysis data. Since observations of groundtruth extinction coefficients are not available in 2D satellite products, we intend to collocate observations from MODIS 2D AOD products (Remer et al., 2005) with CALIOP vertical lidar measurements (Winker et al., 2013) to validate the model.

Acknowledgments. We would like to thank two anonymous reviewers for their thorough reviews and comments, which contributed significantly to the quality of this work.

Author contribution. Conceptualization: S.B., S.S.; Data curation: S.B., D.W.-P.; Investigation: S.B.; Methodology: S.B., D.S., D.W.-P.; Software: S.B.; Visualization: S.B.; Writing—original draft: S.B. All authors approved the final submitted draft.

Competing interest. The authors declare none.

Data availability statement. The data and code that support the findings of this study are openly available at <https://doi.org/10.5281/zenodo.11028725>.

Ethics statement. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Funding statement. S.B., D.W.-P., A.N., and D.S. receive funding from the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No. 860100. D.W.-P. acknowledges funding from NERC project NE/S005390/1 (ACRUISE). S.S. receives funding from the Engineering and Physical Sciences Research Council (EPSRC) via the UK Research and Innovation (UKRI) Centre for Doctoral Training in Application of Artificial Intelligence to the study of Environmental Risks (AI4ER, EP/S022961/1).

Nomenclature

Acronyms

ACIs	aerosol-cloud interactions
AOD	aerosol optical depth
CALIOP	Cloud-Aerosol Lidar with Orthogonal Polarization
CCN	cloud condensation nuclei
ECHAM-HAM	ECMWF Hamburg-Hamburg Aerosol Model
ECMWF	European Centre for Medium-Range Weather Forecasts
ELBO	evidence lower-bound
GP	Gaussian process
ICI	integrated calibration index

MISR	multi-angle imaging spectroradiometer
MODIS	moderate resolution imaging spectroradiometer
SV	Shapley value
VIIRS	visible infrared imaging radiometer suite

Greek letters

α	truncated credible interval size
γ_1	spatiotemporal kernel variance hyperparameter
γ_2	meteorological predictors kernel variance hyperparameter
η	generic notation for the mean parameter of a log-normal distribution
η_i	mean parameter of the log-normal observation model for the i th column
$\hat{\eta}_i$	approximated mean parameter of the log-normal observation model for the i th column
λ	generic notation for wavelength
μ	generic notation for the location parameter of a log-normal distribution
$\boldsymbol{\mu}_z$	mean vector of the variational distribution
$\bar{\boldsymbol{\mu}}(\cdot)$	mean vector of the posterior variational distribution evaluated at (\cdot)
ν	Matérn covariance function order
σ	scale parameter of the AOD log-normal observation model
σ_{ext}	scale parameter of the extinction coefficient log-normal observation model
$\boldsymbol{\Sigma}_z$	covariance matrix of the variational distribution
$\bar{\boldsymbol{\Sigma}}(\cdot)$	covariance of the posterior variational distribution evaluated at (\cdot)
τ	generic notation for the observed AOD
τ_i	observed AOD for the i th column
${}^s\tau_i$	spatially smoothed AOD observation for the i th column
$\boldsymbol{\tau}$	vector of observed AOD for the n columns
φ	prior over extinction coefficient profile
${}^s\varphi$	rescaled prior over extinction coefficient profile
ψ	positive link function warping the GP

Latin letters

b_{ext}	extinction coefficient
d	dimensionality of the vertically resolved covariates vector
\mathcal{D}	dataset
f	Gaussian process
f	generic notation for a realization of the GP at vertically resolved covariates $x h$
\mathbf{f}	realization of the GP overall vertically resolved inputs in the dataset
\mathbf{f}_i	realization of the GP overall vertically resolved inputs from the i th column
h	generic notation for height
$h_i^{(j)}$	altitude of the j th vertically resolved covariates vector for the i th column
H	generic notation for the total height of an atmospheric column
\mathbf{I}_p	identity matrix of size p
k	GP covariance function or kernel
\mathbf{K}_{zz}	covariance matrix evaluated at inducing locations
$\ell(\cdot)$	kernel length scale hyperparameter for variable (\cdot)
L	idealized exponential profile length scale
lat	generic notation for latitude
lon	generic notation for longitude
m	GP mean function
m_i	number of observed altitude levels for the i th column
M	total number of vertically resolved observations across the dataset
n	number of columns observed in the dataset

$N_{\sigma_{\text{ext}}}(\alpha)$	percentage of groundtruth observation falling within the $1 - \alpha$ credible interval
p	number of inducing locations
P	generic notation for pressure
q	variational distribution
S	generic notation for supersaturation
t	generic notation for time
T	generic notation for temperature
\mathbf{u}	inducing variables of the variational distribution
w	weighting function for the idealized exponential profile
\mathbf{z}	inducing locations of the variational distribution
x	generic notation for the vertically resolved covariates vector
$x_i^{(j)}$	observed vertically resolved covariates at altitude $h_i^{(j)}$ for the i th column
\mathbf{x}	vector of all vertically resolved covariates from the dataset
\mathcal{X}	space in which the vertically resolved covariates vector take values

References

- Ackerman AS, Kirkpatrick MP, Stevens DE and Toon OB (2004) The impact of humidity above stratiform clouds on indirect aerosol climate forcing. *Nature*, 432(7020), 1014–1017.
- Albrecht BA (1989) Aerosols, cloud microphysics, and fractional cloudiness. *Science*, 245(4923), 1227–1230.
- Andreae MO (2009) Correlation between cloud condensation nuclei concentration and aerosol optical thickness in remote and polluted regions. *Atmospheric Chemistry and Physics*, 9(2), 543–556.
- Borovitskiy V, Terenin A, Mostowsky P and Deisenroth MP (2020) Matérn Gaussian processes on Riemannian manifolds. *Advances in Neural Information Processing Systems* 33, 12426–12437.
- Camps-Valls G, Verrelst J, Muñoz-Marí J, Laparra V, Mateo-Jimenez F and Gómez-Dans JL (2016) A survey on Gaussian processes for earth-observation data analysis: A comprehensive investigation. *IEEE Geoscience and Remote Sensing Magazine*, 4(2), 58–78.
- Carlaw KS, Lee LA, Reddington CL, Pringle KJ, Rap A, Forster PM, Mann GW, Spracklen DV, Woodhouse MT, Regayre LA and Pierce JR (2013) Large contribution of natural aerosols to uncertainty in indirect forcing. *Nature*, 503(7474), 67–71.
- Carlaw KS and Pringle K (2022) Chapter 4 – Global aerosol properties. In Carlaw KS (ed.), *Aerosols and Climate*. Amsterdam: Elsevier, pp. 101–133. <https://doi.org/10.1016/B978-0-12-819766-0.00011-0>. Available at <https://www.sciencedirect.com/science/article/pii/B9780128197660000110>.
- Chau SL, Bouabid S and Sejdinovic D (2021) Deconditional downscaling with Gaussian processes. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 17813–17825.
- Chen Y-C, Christensen MW, Stephens GL and Seinfeld JH (2014) Satellite-based estimate of global aerosol-cloud radiative forcing by marine warm clouds. *Nature Geoscience*, 7(9), 643–646.
- Christensen MW, Neubauer D, Poulsen CA, Thomas GE, McGarragh GR, Povey AC, Proud SR and Grainger RG (2017) Unveiling aerosol–cloud interactions – Part 1: Cloud contamination in satellite products enhances the aerosol indirect forcing estimate. *Atmospheric Chemistry and Physics*, 17(21), 13151–13164.
- Clarke A and Kapustin V (2010) Hemispheric aerosol vertical profiles: Anthropogenic impacts on optical depth and cloud nuclei. *Science*, 329(5998), 1488–1492.
- Douglas A and L'Ecuyer T (2020) Quantifying cloud adjustments and the radiative forcing due to aerosol–cloud interactions in satellite observations of warm marine clouds. *Atmospheric Chemistry and Physics*, 20(10), 6225–6241.
- Feragen A and Hauberg S (2016) Open problem: Kernel methods on manifolds and metric spaces. What is the probability of a positive definite geodesic exponential kernel? In *Conference on Learning Theory*. Columbia University, New York, New York, USA: PMLR, pp. 1647–1650.
- Ghorbani A and Zou J (2019) Data Shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning*. Long Beach, CA: PMLR, pp. 2242–2251.
- Hoffman MD, Blei DM, Wang C and Paisley J (2013) Stochastic variational inference. *Journal of Machine Learning Research*, 14(40), 1303–1347.
- Holben BN, Eck TF, Slutsker I, Tanré D, Buis JP, Setzer A, Vermote E, Reagan JA, Kaufman YJ, Nakajima T, Lavenu F, Jankowiak I and Smirnov A (1998) AERONET: A federated instrument network and data archive for aerosol characterization. *Remote Sensing of Environment*, 66(1), 1–16.
- Jesson A, Douglas A, Manshausen P, Meinshausen N, Stier P, Gal Y and Shalit U (2022) Scalable sensitivity and uncertainty analysis for causal-effect estimates of continuous-valued interventions. *Advances in Neural Information Processing Systems*, 35, 13892–13907.

- Kahn RA and Gaitley BJ** (2015) An analysis of global aerosol type as retrieved by MISR. *Journal of Geophysical Research (Atmospheres)*, 120(9), 4248–4281.
- Kingma DP and Ba J** (2015) Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*. San Diego, CA: ICLR.
- Kingma DP, Salimans T and Welling M** (2015) Variational dropout and the local reparameterization trick. *Advances in Neural Information Processing Systems*.
- Köhler H** (1936) The nucleus in and the growth of hygroscopic droplets. *Transactions of the Faraday Society*, 32, 1152–1161.
- Laszlo I and Liu HQ** (2016) EPS aerosol optical depth (AOD) algorithm theoretical basis document. *NOAA NESDIS Center for Satellite Application and Research*.
- Law HCL, Sejdinovic D, Cameron E, Tim CDL, Flaxman S, Battle K and Fukumizu K** (2018) Variational learning on aggregate outputs with Gaussian processes. *Advances in Neural Information Processing Systems*.
- Law HC, Zhao P, Chan LS, Huang J and Sejdinovic D** (2019) Hyperparameter learning via distributional transfer. *Advances in Neural Information Processing Systems*.
- Leibfried F, Dutordoir V, John ST and Durrande N** (2020) A tutorial on sparse Gaussian processes and variational inference. arXiv preprint [arXiv:2012.13962](https://arxiv.org/abs/2012.13962).
- Leinonen J, Guillaume A and Yuan T** (2019) Reconstruction of cloud vertical structure with a generative adversarial network. *Geophysical Research Letters*, 46, 7035–7044.
- Levy RC, Remer LA, Mattoo S, Vermote EF and Kaufman YJ** (2007) Second-generation operational algorithm: Retrieval of aerosol properties over land from inversion of moderate resolution imaging Spectroradiometer spectral reflectance. *Journal of Geophysical Research: Atmospheres*, 112, D13211.
- Li C, Li J, Dubovik O, Zeng Z-C and Yung YL** (2020) Impact of aerosol vertical distribution on aerosol optical depth retrieval from passive satellite sensors. *Remote Sensing*, 12(9), 1524.
- Lundberg SM, Erion GG and Lee S-I** (2018) Consistent individualized feature attribution for tree ensembles. arXiv preprint [arXiv:1802.03888](https://arxiv.org/abs/1802.03888).
- Lundberg SM and Lee S-I** (2017) A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* 30, 4765–4774.
- Masson-Delmotte V, Zhai P, Pirani A, Connors SL, Péan C, Berger S, Caud N, Chen Y, Goldfarb L, Gomis MI, Huang M, Leitzell K, Lonnoy E, Matthews JBR, Maycock TK, Waterfield T, Yelekçi O, Yu R and Zhou B** (eds.) (2021) *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press.
- Matthews AGG, Hensman J, Turner R and Ghahramani Z** (2016) On sparse variational methods and the Kullback–Leibler divergence between stochastic processes. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*. Cadiz, Spain. JMLR.
- McCormick RA and Ludwig JH** (1967) Climate modification by atmospheric aerosols. *Science*, 156(3780), 1358–1359.
- Mielonen T, Levy RC, Aaltonen V, Komppula M, de Leeuw G, Huttunen J, Lihavainen H, Kolmonen P, Lehtinen KEJ and Arola A** (2011) Evaluating the assumptions of surface reflectance and aerosol type selection within the MODIS aerosol retrieval over land: The problem of dust type selection. *Atmospheric Measurement Techniques*, 4(2), 201–214.
- Muñoz-Sabater J, Dutra E, Agustí-Panareda A, Albergel C, Arduini G, Balsamo G, Boussetta S, Choulga M, Harrigan S, Hersbach H, Martens B, Miralles DG, Piles M, Rodríguez-Fernández NJ, Zsoter E, Buontempo C and Thépaut J-N** (2021) ERA5-Land: A state-of-the-art global reanalysis dataset for land applications.
- Nakajima T, Higurashi A, Kawamoto K and Penner JE** (2001) A possible correlation between satellite-derived cloud and aerosol microphysical parameters. *Geophysical Research Letters*, 28(7), 1171–1174.
- O'Neill NT, Ignatov A, Holben BN and Eck TF** (2000) The lognormal distribution as a reference for reporting aerosol optical depth statistics; empirical tests using multi-year, multi-site AERONET Sunphotometer data. *Geophysical Research Letters*, 27(20), 3333–3336.
- Painemal D, Chang F-L, Ferrare R, Burton S, Li Z, Smith WL, Minnis P, Feng Y and Clayton M** (2020) Reducing uncertainties in satellite estimates of aerosol–cloud interactions over the subtropical ocean by integrating vertically resolved aerosol observations. *Atmospheric Chemistry and Physics*, 20(12), 7167–7177.
- Rasmussen CE and Williams CKI** (2005) *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press.
- Remer LA, Kaufman YJ, Tanré D, Mattoo S, Chu DA, Martins JV, Li R-R, Ichoku C, Levy RC, Kleidman RG, Eck TF, Vermote E and Holben BN** (2005) The MODIS aerosol algorithm, products, and validation. *Journal of the Atmospheric Sciences*, 62(4), 947–973.
- Shapley LS** (1953) A value for n-person games. *Contributions to the Theory of Games*.
- Small JD, Chuang PY, Feingold G and Jiang H** (2009) Can aerosol decrease cloud lifetime? *Geophysical Research Letters*, 36(16).
- Spracklen DV, Carslaw KS, Pöschl U, Rap A and Forster PM** (2011) Global cloud condensation nuclei influenced by carbonaceous combustion aerosol. *Atmospheric Chemistry and Physics*, 11(17), 9067–9087.
- Stein ML** (1999) *Interpolation of Spatial Data: Some Theory for Kriging*. New York: Springer Science & Business Media.
- Stier P** (2016) Limitations of passive remote sensing to constrain global cloud condensation nuclei. *Atmospheric Chemistry and Physics*, 16(10), 6595–6607.

Stier P, Feichter J, Kinne S, Kloster S, Vignati E, Wilson J, Ganzeveld L, Tegen I, Werner M, Balkanski Y, Schulz M, Boucher O, Minikin A and Petzold A (2005) The aerosol-climate model ECHAM5-HAM. *Atmospheric Chemistry and Physics*, 5(4), 1125–1156.

Stier P, Seinfeld JH, Kinne S and Boucher O (2007) Aerosol absorption and radiative forcing. *Atmospheric Chemistry and Physics*, 7(19), 5237–5261.

Štrumbelj E and Kononenko I (2014) Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41, 647–665.

Tanaka Y, Tanaka T, Iwata T, Kurashima T, Okawa M, Akagi Y and Toda H (2019) Spatially aggregated Gaussian processes with multivariate areal outputs. *Advances in Neural Information Processing Systems*.

Tanskanen AKV and Longi K (2020) Non-linearities in Gaussian processes with integral observations. *IEEE international Workshop on Machine Learning for Signal*.

Titsias MK (2009) Variational learning of inducing variables in sparse Gaussian processes. In *AISTATS*. Clearwater Beach, FL: JMLR.

Twomey S (1974) Pollution and the planetary albedo. *Atmospheric Environment*, 8(12), 1251–1256.

Twomey S (1977) The influence of pollution on the shortwave albedo of clouds. *Journal of Atmospheric Sciences*, 34, 1149–1152.

Watson-Parris D, Bellouin N, Deaconu LT, Schutgens NAJ, Yoshioka M, Regayre LA, Pringle KJ, Johnson JS, Smith CJ, Carslaw KS and Stier P (2020) Constraining uncertainty in aerosol direct forcing. *Geophysical Research Letters*, 47(9), e2020GL087141.

Winker DM, Tackett JL, Getzewich BJ, Liu Z, Vaughan MA and Rogers RR (2013) The global 3-D distribution of tropospheric aerosols as characterized by CALIOP. *Atmospheric Chemistry and Physics*, 13(6), 3345–3361.

Wu Y, De Graaf M and Menenti M (2016) The sensitivity of AOD retrieval to aerosol type and vertical distribution over land with MODIS data. *Remote Sensing*, 8(9), 765.

Wu Y, de Graaf M and Menenti M (2017) The impact of aerosol vertical distribution on aerosol optical depth retrieval using Calipso and Modis data: Case study over dust and smoke regions. *Journal of Geophysical Research: Atmospheres* 122(16), 8801–8815.

Yousefi F, Smith MT and Álvarez MA (2019) Multi-task learning for aggregated data using Gaussian processes. *Advances in Neural Information Processing Systems*.

Zhang Y, Charoenphakdee N, Wu Z and Sugiyama M (2020) Learning from aggregate observations. *Advances in Neural Information Processing Systems*.

Zhang K, O'Donnell D, Kazil J, Stier P, Kinne S, Lohmann U, Ferrachat S, Croft B, Quaas J, Wan H, Rast S and Feichter J (2012) The global aerosol-climate model ECHAM-HAM, version 2: Sensitivity to improvements in process representations. *Atmospheric Chemistry and Physics*, 12(19), 8911–8949.

Zhou Z-H (2017) A brief introduction to weakly supervised learning. *National Science Review*, 5(1), 44–53.

A. Appendix A: Matérn covariance

The Matérn covariances are a class of stationary covariance functions widely used in spatial statistics. The Matérn- ν covariance with unit variance between two points $x, x' \in \mathbb{R}$ is given by

$$C_\nu(|x - x'|) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{|x - x'|}{\ell} \right)^\nu K_\nu \left(\sqrt{2\nu} \frac{|x - x'|}{\ell} \right), \tag{A-1}$$

where Γ is the gamma function, K_ν is the modified Bessel function, and $\ell > 0$ is a length scale hyperparameter.

The covariance function expression considerably simplifies for $\nu = p + 1/2$ where $p \in \mathbb{N}$. For example, for $\nu = 1/2$ ($p = 0$) and $\nu = 3/2$ ($p = 1$) we have

$$C_{1/2}(|x - x'|) = \exp\left(-\frac{|x - x'|}{\ell}\right), \tag{A-2}$$

$$C_{3/2}(|x - x'|) = \left(1 + \sqrt{3} \frac{|x - x'|}{\ell}\right) \exp\left(-\sqrt{3} \frac{|x - x'|}{\ell}\right). \tag{A-3}$$

When $x, x' \in \mathbb{R}^d$, the distance $|x - x'|$ can be substituted by the norm $\|x - x'\| = \sqrt{\sum_{i=1}^d (x_i - x'_i)^2}$. The covariance is called an *automatic relevance determination* (ARD) kernel when each dimension has its own independent length scale parameter $\ell_i > 0$. For example, the Matérn-1/2 and Matérn-3/2 ARD kernel write

$$\|x - x'\|_\ell = \sqrt{\sum_{i=1}^d \frac{(x_i - x'_i)^2}{\ell_i}}, \tag{A-4}$$

$$C_{1/2}(\|x - x'\|_\ell) = \exp(-\|x - x'\|_\ell), \tag{A-5}$$

$$C_{3/2}(\|x - x'\|_\ell) = \left(1 + \sqrt{3} \|x - x'\|_\ell\right) \exp\left(-\sqrt{3} \|x - x'\|_\ell\right). \tag{A-6}$$

When a Matérn- $p+1/2$ covariance function is used as a kernel for a GP, draws from the GP are p times continuously differentiable (with convention that 0 times means simply continuous).

B. Appendix B: Comparison with tensor product covariance structure

A standard choice in the design of covariance structures is to assume a tensor product form of the covariance with an independent correlation over each input dimension, that is, $k(x, x') = \prod_i k(x_i, x'_i)$. We compare our choice of covariance structure from Section 3.2.2 to the product covariance function given by

$$k(x|h, x'|h') = \gamma_2 C_{3/2}(|t - t'|) C_{3/2} \left(\text{Hav} \left(\begin{bmatrix} \text{lat} \\ \text{lon} \end{bmatrix}, \begin{bmatrix} \text{lat}' \\ \text{lon}' \end{bmatrix} \right) \right) \tag{B-1}$$

$$\times C_{1/2}(|T - T'|) C_{1/2}(|P - P'|) C_{1/2}(|RH - RH'|) C_{1/2}(|\omega - \omega'|), \tag{B-2}$$

where we maintain the same Matérn covariances for the sake of comparison. This product structure introduces a global variance parameter γ_2 and independent length scale parameters $\ell_t, \ell_{\text{lat,lon}}, \ell_T, \ell_P, \ell_{RH}, \ell_\omega$ for each input variable, resulting in seven hyperparameters that need to be tuned.

In contrast, the covariance structure we propose in Section 3.2.2 introduces two variance parameters γ_1, γ_2 and an equal number of length scale parameters for each input variable (through the ARD Matérn kernel on meteorological predictors), resulting in eight hyperparameters that need to be tuned. However, because we fix $\gamma_2 = 1$ in our model, the number of hyperparameters that effectively require tuning is 7.

Table B1 reports the hyperparameters values for both the product covariance and our covariance, after tuning them to maximize the ELBO objective as described in Section 2.3.3. We find that the product covariance structure leads to larger length scale parameters for the meteorological predictors. This means that, in general, predictions with the product covariance structure should have less sensitivity to changes in meteorological predictors.

Using the tuned models, we then perform inference and predict the ECHAM-HAM extinction coefficients over the dataset with both models following the procedure outlined in Section 2.3. Table B2 reports the ECHAM-HAM extinction prediction scores for the product covariance and our covariance. It shows the proposed covariance provides substantial improvement in deterministic metrics compared to the tensor product covariance. Regarding probabilistic scores, the ELBO and Calib95 are less sensitive to the choice of covariance structure. We find that the ICI is improved over the entire column with a product covariance structure, but

Table B1. Hyperparameter values for our choice of covariance and a tensor product covariance structure

Covariance	γ_1	γ_2	ℓ_t	$\ell_{\text{lat,lon}}$	ℓ_T	ℓ_P	ℓ_{RH}	ℓ_ω
Ours	0.26 ± 0.02	1.0^\dagger	7.6 ± 0.2	0.85 ± 0.06	2.5 ± 0.1	3.4 ± 0.1	3.4 ± 0.1	6.6 ± 0.1
Product	–	0.92 ± 0.1	8.8 ± 0.1	1.2 ± 0.1	5.6 ± 0.1	6.2 ± 0.1	5.8 ± 0.1	8.3 ± 0.1

Note. Models are tuned for standardized input variables following the procedure described in Section 2.3.3; hyperparameters are tuned using 10 different initialization seeds; we report 1 standard deviation; † indicates the value of the parameter is fixed in the model and not tuned against data.

Table B2. Scores of the VAExtGP with our choice of covariance and a product covariance structure for the task of predicting ECHAM-HAM extinction profiles

Region	Covariance	RMSE (10^{-5})	MAE (10^{-6})	Corr	Bias (10^{-6})	Bias98 (10^{-5})
Entire column	Product	3.65 ± 0.03	5.92 ± 0.09	0.61 ± 0.01	-0.54 ± 0.01	-1.15 ± 0.01
	Ours	3.21 ± 0.04	5.29 ± 0.06	0.73 ± 0.01	-0.005 ± 0.161	-0.482 ± 0.212
Boundary layer	Product	6.73 ± 0.05	16.5 ± 0.1	0.59 ± 0.01	-3.46 ± 0.22	-5.64 ± 0.28
	Ours	5.91 ± 0.07	15.3 ± 0.1	0.72 ± 0.06	-0.404 ± 0.71	-4.53 ± 0.44

Region	Covariance	ELBO	Calib95 (%)	ICI (10^{-2})
Entire column	Product	13.0 ± 0.1	93.9 ± 0.1	4.45 ± 0.15
	Ours	13.0 ± 0.1	93.5 ± 0.1	6.60 ± 0.45
Boundary layer	Product	10.6 ± 0.1	98.3 ± 0.1	7.16 ± 0.20
	Ours	10.7 ± 0.1	98.7 ± 0.1	6.26 ± 0.34

Note. “Entire column” means scores are computed and averaged for every altitude level; “Boundary layer” means scores are computed and averaged for altitude levels of the boundary level only (<2 km); the best scores for each metric and region are highlighted in bold; we report 1 standard deviation.

degraded within the boundary layer. These results suggest that separating the contribution of spatiotemporal and meteorological predictors in the covariance benefits extinction predictive performance.

C. Appendix C: Additional experimental details

C.1. Ablation study: details on relative change computation

To evaluate the relative change in scores of the ablated model against *VAExtGP*, we compute

$$\% \text{Relative change} = \frac{\text{Score } VAExtGP - \text{Score ablated model}}{\text{Score } VAExtGP} \times 100, \quad (\text{C-1})$$

where a positive percentage indicates improvement with respect to *VAExtGP* and a negative percentage indicates deterioration with respect to *VAExtGP*. However, this is only a useful indicator of change in performance across scores if all scores are improved when smaller. This is true for RMSE, MAE, and ICI, but not for the other metrics used. Therefore, we take the following transformations reported in Table C1 to ensure that smaller values correspond to an improvement in score.

Table C1. Evaluation metrics and transformed evaluation metrics are used to compute the relative change in scores for ablation experiments

Metric	Best when	Transformed metric	Best when
RMSE	Lower	RMSE	Lower
MAE	Lower	MAE	Lower
Corr	Higher	$-1 \times \text{Corr}$	Lower
Bias	Close to 0	$ \text{Bias} $	Lower
Bias98	Close to 0	$ \text{Bias}_{98} $	Lower
ELBO	Higher	$-1 \times \text{ELBO}$	Lower
Calib95	Close to 95%	$ \text{Calib}_{95} - 95\% $	Lower
ICI	Lower	ICI	Lower

C.2. Additional sliced plots

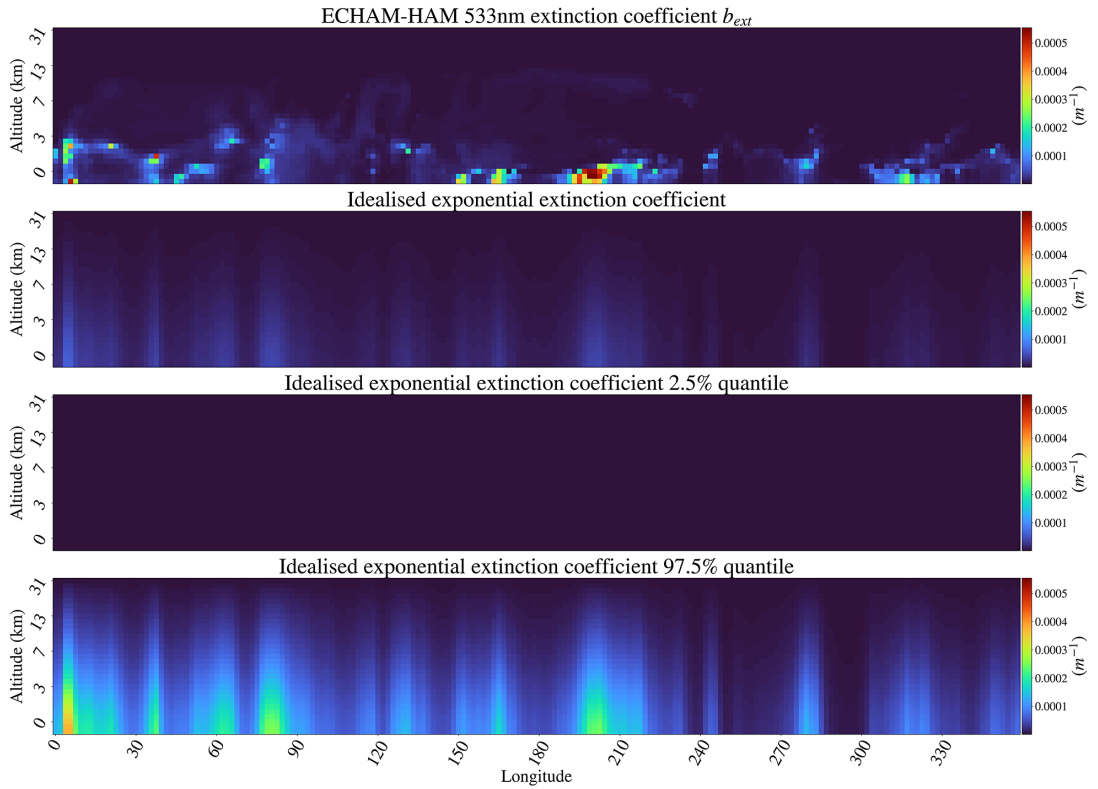


Figure C1. Vertical slices at latitude 51.29° of groundtruth extinction coefficient, idealized exponential extinction coefficient, 2.5% and 97.5% quantiles of the idealized exponential extinction coefficient distribution.

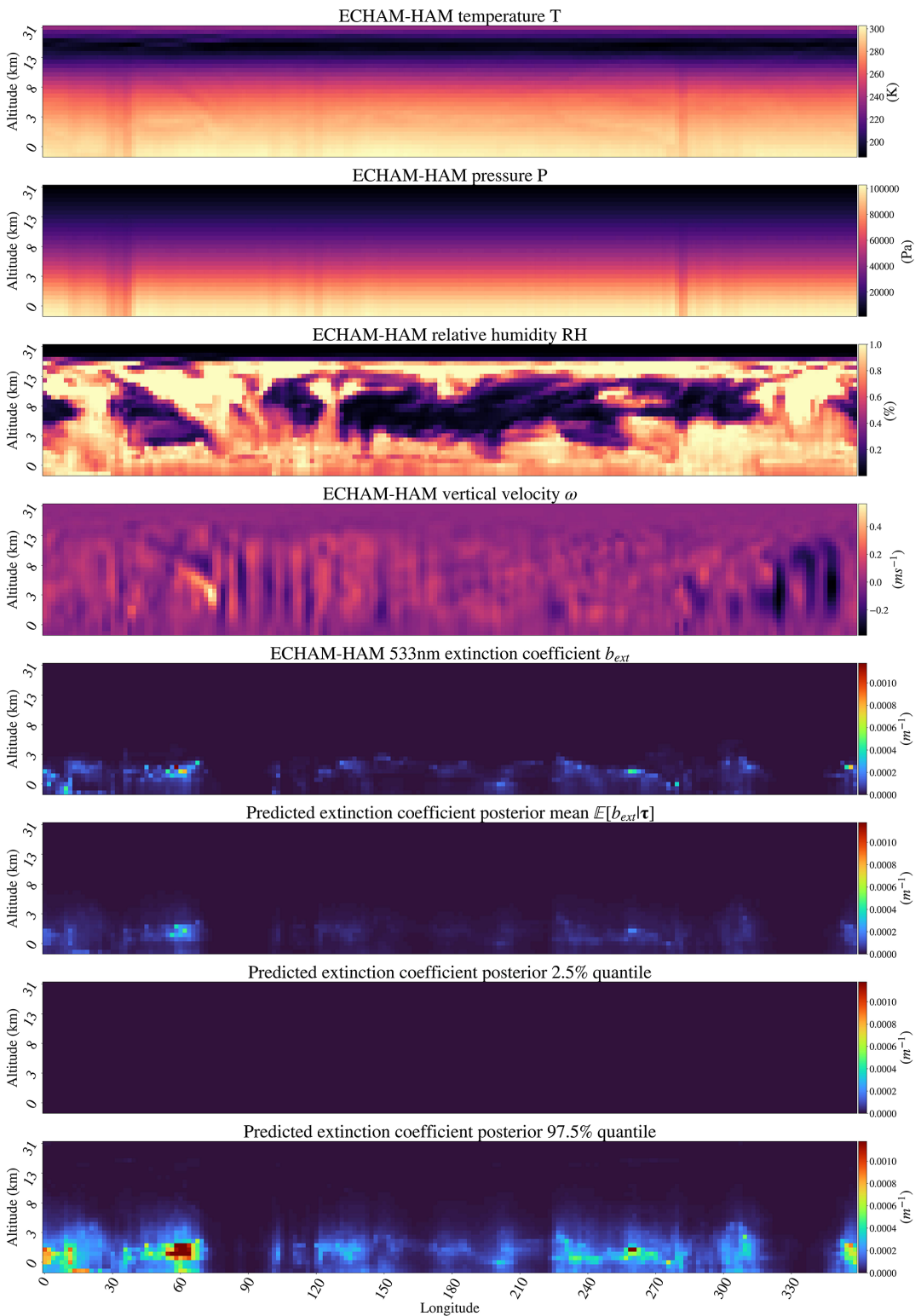


Figure C2. Vertical slices at latitude -0.93° of meteorological predictors (T, P, RH, ω), groundtruth extinction coefficient, predicted extinction coefficient posterior mean, 2.5% and 97.5% quantiles of the predicted extinction coefficient posterior distribution.

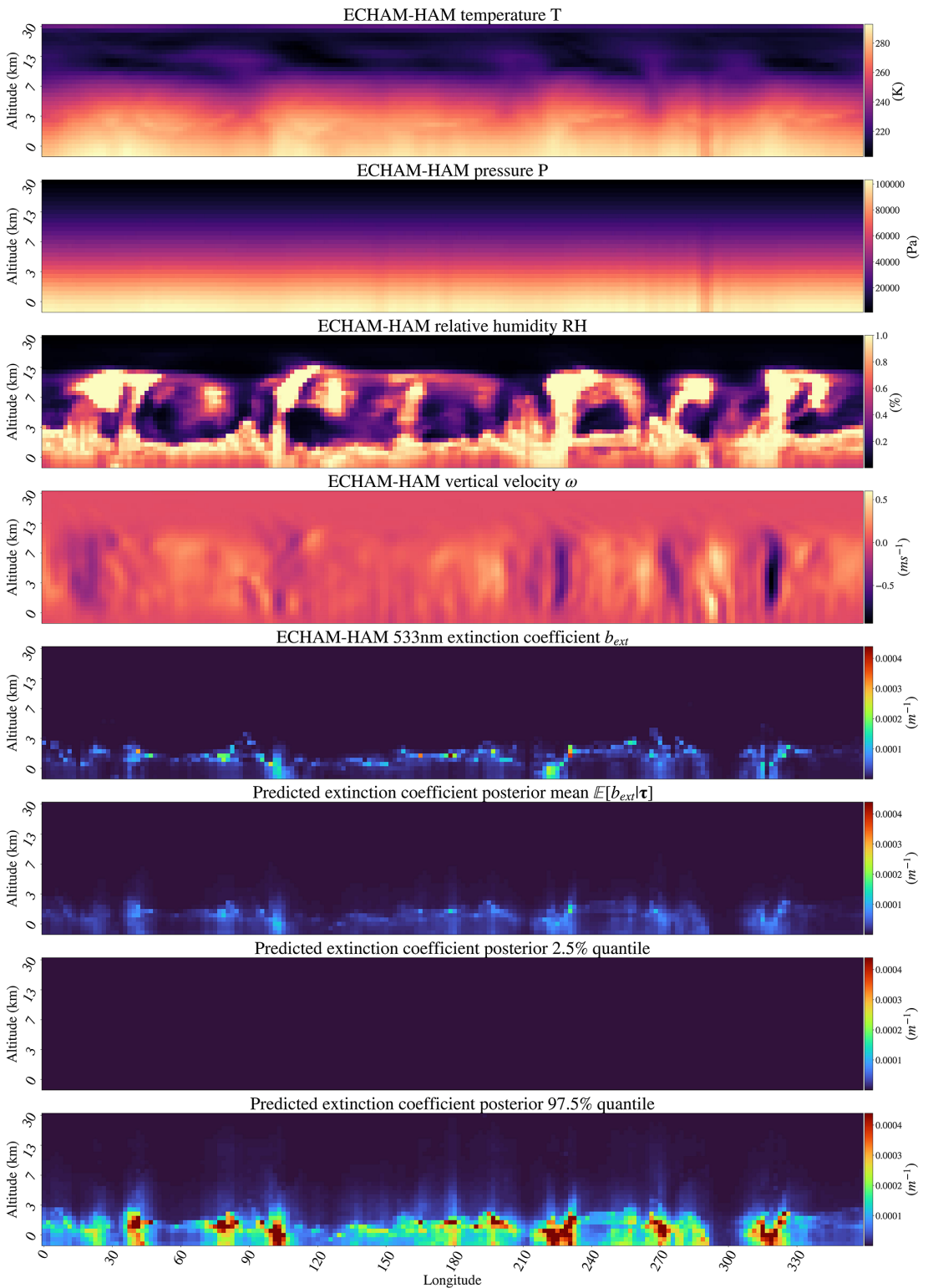


Figure C3. Vertical slices at latitude -38.2° of meteorological predictors (T, P, RH, ω), groundtruth extinction coefficient, predicted extinction coefficient posterior mean, 2.5% and 97.5% quantiles of the predicted extinction coefficient posterior distribution.