Invertible Neural Networks for Probabilistic Aerosol Optical Depth Retrieval

Paolo Pelucchi, Jorge Vicent Servera, Philip Stier, Gustau Camps-Valls *Fellow, IEEE*

***Abstract*—Satellite remote sensing is the primary source of global aerosol observations, providing essential data for understanding aerosol-climate interactions and constraining global climate models. To solve the inverse problem at the heart of the retrieval process, traditional algorithms must make simplifications and often cannot quantify uncertainty. In this study, we explore the use of Invertible Neural Networks (INNs) for retrieving aerosol optical depth (AOD) from spectral top-of-atmosphere reflectance. INNs can handle the inherent uncertainty of underdetermined inverse problems. They model the forward and inverse processes simultaneously, while learning additional random latent variables used to recover full non-parametric posterior distributions for the inverse predictions. We develop location-specific INNs for MODIS sensor data, training on synthetic datasets generated by combining atmospheric reflectance from MODIS Dark Target (DT) look-up tables and surface reflectance from a MODIS bidirectional reflectance product. The INNs successfully emulate the forward problem, and achieve accurate AOD inversion results on synthetic test sets (RMSE $\approx$ 0.05). The posterior distributions obtained are reliable (mean absolute calibration error $\approx$ 2.5%), efficiently providing informative predictive uncertainty estimates. Additionally, the INNs' invertible architecture is found to promote physically consistent predictions and uncertainties. To further validate them in a real-world context, the INNs are applied to MODIS L1B reflectance observations to produce full-resolution AOD estimates with pixel-level uncertainties. The retrievals are compared to collocated ground measurements from the Aeronet network. The INNs obtain good accuracy in all tested locations in line with the operational DT AOD product (RMSE $\approx$ 0.1, 74% within DT expected error bounds). The INNs are also able to retrieve AOD over bright surfaces where DT cannot be applied. Despite uncovered limitations out-of-distribution, the INNs show consistent skill in target domains across diverse land surfaces. The INNs' unique modelling and uncertainty quantification features have the potential to enhance aerosol and climate studies in various real-world contexts.**

***Index Terms*—Remote sensing, satellite retrievals, aerosol optical depth, invertible neural networks, uncertainty estimation.**

I. Introduction

UNDERSTANDING the properties of atmospheric aerosols, microscopic particles suspended in the air, is crucial for many scientific and societal applications. These particles significantly impact air quality, climate, and the Earth's energy budget [1, 2]. Still, their direct measurement across vast spatial and temporal scales remains challenging. Satellite remote sensing is the primary source of observations about aerosols on a global scale, providing essential data over large areas and with regular revisit times. However, retrieving accurate aerosol properties from satellite observations constitutes a complex ill-posed inverse problem [3], i.e., a problem in which multiple physical causes (in this case aerosols and surface reflectance) must be recovered from observations of their effects (backscattered radiation). This results in an underdetermined problem where multiple solutions are potentially compatible with any given observation. Hence, a probabilistic description is essential to capture the problem's inherent uncertainty.

The forward problem can be simulated by radiative transfer models (RTMs). Based on well-known physical laws, RTMs can calculate top-of-atmosphere (TOA) observed reflectances at different wavelengths for prescribed atmospheric and surface conditions. Traditional aerosol retrieval algorithms make use of information from RTMs to address the inverse problem. Firstly, the algorithm must determine the spectral surface reflectances from the the satellite-observed TOA signal, separating them from the atmospheric components which contain information about the aerosols. An estimate of the surface reflectance may be obtained in a channel transparent to most aerosols, and the surface reflectance in other channels derived based on empirical relationships [4]. Such relationships can be further parameterised to account for the effects of e.g. vegetation conditions, scattering angle and urban surfaces [5, 6]. Another possibility is to construct and use observational surface databases to estimate surface reflectance based on geolocation and time of year; this approach is often combined with spectral reflectance relationships for improved accuracy [7, 8].

Fixing the surface reflectance necessarily determines the atmospheric reflectance. The atmospheric reflectances contain information about the gases and aerosols present in the atmospheric column. But while for gas retrievals only the amount must be determined, aerosols generally occur as a mixture of lognormal modes with varying composition and size, which the algorithm must also determine in addition to the aerosol amount (measured as aerosol optical depth, AOD, at $550\,\mathrm{nm}$). Multi-angle, multi-polarisation instruments have the ability to recover such parameters [9]. However, for most single-angle sensors there is little information content on composition and size distribution in TOA measurements, so retrieval algorithms tend to refer back to the prior, selected from a handful of possible aerosol models. In fact, the atmospheric reflectances

PP, JVS and GCV are with the Image Processing Laboratory, Universitat de València, Valencia, Spain (email: paolo.pelucchi@uv.es).

PS is with the Department of Physics, University of Oxford, Oxford, UK.

Manuscript submitted December 6, 2024. Work supported the Marie Curie ITN iMIRACLI project 'innovative MachIne leaRning to constrain Aerosol-cloud CLimate Impacts' (860100) and the by the ERC under the ERC-SyG-2019 USMILE project 'Understanding and Modelling the Earth System with Machine Learning' (grant agreement 855187).

are compared with a set of pre-computed RTM reflectances, corresponding to different aerosol models and AOD levels, and the closest match is selected as the retrieval solution. Such method is known as the look-up table (LUT) approach [10], and is used in many operational aerosol retrieval algorithms [11, 12]. This simplified approach saves on computational cost and provides a single solution to the inverse problem, but potentially compromises accuracy and limits the ability to quantify uncertainties. Some statistical inversion techniques, such as optimal estimation [3, 13] and the multi-term least square method [14], treat uncertainties explicitly by defining covariance matrices for various sources of error. The retrieval process involves repeatedly utilizing the computationally expensive forward RTMs to iteratively arrive at a solution, which hinders practicality. The method also requires making some linearity and Gaussianity assumptions [15].

Machine learning (ML) techniques have found growing use in aerosol remote sensing problems due to their ability to learn from data and reproduce complex relationships. They can be used to enhance traditional physics-based algorithms, increasing accuracy where assumptions or simplifications are generally needed. ML methods (neural networks in particular) have been used, for example, to emulate RTMs and replace LUT interpolation [16], to derive improved surface reflectance relationships [17], and to detect dust aerosol plumes [18]. Another common application is that of post-process bias correction, in which a ML model is trained to make satellite AOD retrievals obtained by a traditional algorithm closer to those of a more accurate validation product such as Aeronet [19, 20]. Fully ML-based AOD retrieval algorithms have also been proposed, typically trained using collocated datasets of satellite observations and ground-based AOD measurements as ground truth [21]. These models are often supplemented with additional auxiliary variables relating to meteorological conditions, land cover or vegetation to aid retrieval [22, 23]. Neural networks have most commonly been employed in this context [24, 25], but more recently decision tree-based methods like random forests and gradient boosting algorithms have been explored [23, 26]. Ensemble techniques are sometimes used, where predictions from multiple models are averaged to reduce random errors and provide a way of quantifying uncertainty [22, 27]. To further reduce bias, some approaches develop separate models for distinct aerosol conditions, such as high and low AOD, and then integrate their predictions intelligently in a two-stage process [26, 28]. ML-based AOD retrieval models have demonstrated their ability to achieve accurate results, but they are sometimes criticised for lacking physical interpretability [21].

Recently, Invertible Neural Networks (INNs) have emerged as a promising machine learning approach for tackling inverse problems [29]. Unlike conventional physics- and ML-based methods, INNs model both the forward and inverse branches of the problem simultaneously. This unique capability allows them to learn the intricate relationships between causes and effects (state variables and observations) while capturing the associated conditional distributions. INNs efficiently provide non-parametric posterior distributions for the retrieved parameters, offering a valuable tool for uncertainty quantification. INNs have succeeded in some idealized low-dimensional problems and simpler scientific retrievals [29]. They have also been used for aerosol properties retrieval from in situ light scattering measurements [30]. However, applying INNs to satellite aerosol retrievals poses unique challenges. The high variability of surface reflectance signals and the often faint aerosol signature in TOA measurements significantly complicate the inverse problem.

This study addresses these challenges by investigating the application of INNs for retrieving AOD with corresponding uncertainty estimates from TOA reflectance in four MODIS channels.

II. Proposed Methodology

A. Problem setting

In a forward-inverse problem, there is a set of variables $\boldsymbol{x}$ that describe the state of the system of interest, but via observation we only have access to a different set of variables $\boldsymbol{y}$. The forward process, deterministic in nature, describes what observations $\boldsymbol{y}$ result from a particular state $\boldsymbol{x}$ and can be modeled as $\boldsymbol{y} = F(\boldsymbol{x})$ thanks to the knowledge of the physics of the system. Solving the inverse problem, i.e., determining what state $\boldsymbol{x}$ produced a specific observation $\boldsymbol{y}$, is not as straightforward. In fact, the observations generally do not contain enough information to uniquely determine the corresponding state: the problem is ill-posed, and different $\boldsymbol{x}$ are compatible with a single $\boldsymbol{y}$. To recover the full space of compatible solutions, we want to model the inverse process as a conditional probability $p(\boldsymbol{x}|\boldsymbol{y})$. A possible approach is to extend the observations $\boldsymbol{y}$ with additional latent variables $\boldsymbol{z}$ that allow to disentangle the possible $\boldsymbol{x}$, see Fig. 1. This is the idea underlying INNs, introduced in [29], which model the forward and inverse problems simultaneously with a bijective mapping between $\boldsymbol{x}$ on one side and $[\boldsymbol{y}, \boldsymbol{z}]$ on the other. Due to the specific loss function used for training, the INNs automatically push so-called latent Gaussian random variables $\boldsymbol{z}$ to encode information about $\boldsymbol{x}$ not present in $\boldsymbol{y}$. Then, executing the inverse pass with a fixed $\boldsymbol{y}$ and resampling $\boldsymbol{z}$ allows to recover the solutions $\boldsymbol{x} \sim p(\boldsymbol{x}|\boldsymbol{y})$. Next, we describe INNs in more detail following [29].

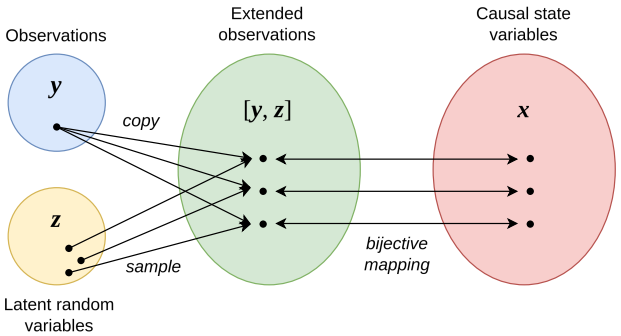

Fig. 1. Invertible neural network idea. The addition of latent random variables $\boldsymbol{z}$ that encode extra information about $\boldsymbol{x}$ allows to extend $\boldsymbol{y}$ and transform the inverse problem from a one-to-many to a bijective mapping. Resampling $\boldsymbol{z}$ while keeping $\boldsymbol{y}$ fixed recovers the possible solutions $\boldsymbol{x} \sim p(\boldsymbol{x}|\boldsymbol{y})$.

B. Invertible Neural Networks

INNs are based on normalizing flows (NFs), a framework recently formalized in machine learning for generative modelling [31]. NFs learn to transform any complex input distribution into a Gaussian distribution by a progressive sequence of invertible and differentiable operations. As the entire process remains invertible, the NF can then be used for density estimation and sampling of the original distribution by working in the Gaussian space and applying the inverse transformations. NFs and similar approaches have found application in many Earth and remote sensing data problems [32]. In INNs, the density transformation idea serves to construct the latent random variables $\boldsymbol{z}$, but the model also simultaneously learns the forward mapping from $\boldsymbol{x}$ to $\boldsymbol{y}$.

The invertible transformation that forms the basic building block of an INN is the affine coupling layer, introduced in [33]. For a given input vector $\boldsymbol{u}$, the affine coupling blocks randomly divide it into two halves $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$, and apply an affine transformation (element-wise multiplication $\odot$ and addition) to each half, giving output vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, which are concatenated and passed on as input to the following block. The forward expressions (1, 2) are easily invertible (3, 4):

$$\boldsymbol{v}_1 = \boldsymbol{u}_1 \odot \exp(s_2(\boldsymbol{u}_2)) + t_2(\boldsymbol{u}_2) \tag{1}$$

$$\boldsymbol{v}_2 = \boldsymbol{u}_2 \odot \exp(s_1(\boldsymbol{v}_1)) + t_1(\boldsymbol{v}_1) \tag{2}$$

$$\boldsymbol{u}_2 = (\boldsymbol{v}_2 - t_1(\boldsymbol{v}_1)) \odot \exp(-s_1(\boldsymbol{v}_1)) \tag{3}$$

$$\boldsymbol{u}_1 = (\boldsymbol{v}_1 - t_2(\boldsymbol{u}_2)) \odot \exp(-s_2(\boldsymbol{u}_2)). \tag{4}$$

The functions s_i and t_i can be arbitrarily complex and do not need to be invertible themselves, as they are always executed in the same direction. Typically, fully-connected neural networks are used and referred to as subnetworks in this context. Their weights are learned automatically during training. The INN architecture requires equal input and output dimensions; zero-padding can be added to balance $\boldsymbol{x}$ and $[\boldsymbol{y}, \boldsymbol{z}]$ if needed.

As mentioned, the $\boldsymbol{y}$ side is extended with a number of dimensions $\boldsymbol{z} \in \mathbb{R}^K$, latent random variables drawn from a standard normal distribution as $\boldsymbol{z} \sim p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_K)$. This allows the inverse problem to be reformulated as a deterministic function of $\boldsymbol{y}$ and $\boldsymbol{z}$: $\boldsymbol{x} = g(\boldsymbol{y}, \boldsymbol{z}; w)$, where w generically refers to function parameters. Then, the INN models the bijective mapping

$$[\boldsymbol{y}, \boldsymbol{z}] = f(\boldsymbol{x}; w) = [f_y(\boldsymbol{y}; w), f_z(\boldsymbol{z}; w)] = g^{-1}(\boldsymbol{x}; w), \tag{5}$$

where $f_y(\boldsymbol{y}; w)$ approximates the forward model $F(\boldsymbol{x})$.

The INNs are trained with a loss function composed of four different terms (Fig. 2):

$$\mathcal{L}_y = \sum\nolimits_i \|\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i\|^2 \tag{6}$$

$$\mathcal{L}_z = \text{MMD}\left(p(\hat{\boldsymbol{y}}, \hat{\boldsymbol{z}}),\ p(\boldsymbol{y})p(\boldsymbol{z})\right) \tag{7}$$

$$\mathcal{L}_x = \text{MMD}\left(p(\hat{\boldsymbol{x}}),\ p(\boldsymbol{x})\right) \tag{8}$$

$$\mathcal{L}_r = \sum\nolimits_i \|g(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{z}}_i; w) - \boldsymbol{x}_i\|^2 \tag{9}$$

where a hat denotes model-generated values. Each term has a specific purpose. The $\mathcal{L}_y$ loss is a standard supervised loss on the forward pass, penalizing deviations between true $\boldsymbol{y}$ and INN predictions $\hat{\boldsymbol{y}}$. The $\mathcal{L}_z$ and $\mathcal{L}_x$ losses use maximum mean discrepancy (MMD), a kernel-based measure of the distance between two distributions evaluated from samples from those distributions [34]. The $\mathcal{L}_z$ loss is pivotal for deriving meaningful latent variables $\boldsymbol{z}$. It minimizes the distance between the joint distribution of $[\hat{\boldsymbol{y}}, \hat{\boldsymbol{z}}]$ predictions from the forward pass and the product of the marginal distributions of true $\boldsymbol{y}$ and sampled Gaussian $\boldsymbol{z}$. This formulation serves a dual purpose: ensuring $\boldsymbol{z}$ adheres to a standard normal distribution and encouraging independence between $\boldsymbol{y}$ and $\boldsymbol{z}$, enabling $\boldsymbol{z}$ to encode unique information. The $\mathcal{L}_x$ loss aligns the backward predictions' distribution with the training dataset's prior distribution, expediting convergence without altering the optimum [29]. Finally, the $\mathcal{L}_r$ reconstruction loss, also used in [30], is added to the previous three described in [29]. It aims to guarantee robustness by verifying that executing the forward and inverse passes returns the original inputs. The total loss is obtained by a weighted sum with coefficients $\gamma_{[y,z,x,r]}$. The INN's weights are updated using losses from forward and inverse passes at each training step.

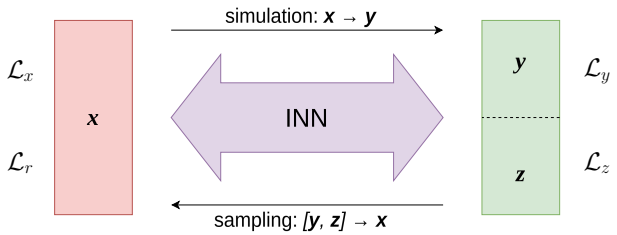

Fig. 2. Invertible neural network implementation. INNs learn latent $\boldsymbol{z}$ and model forward and inverse at once by minimizing a four-term loss. It accounts for error minimization of the forward pass ($\mathcal{L}_y$) and the inverse reconstruction ($\mathcal{L}_r$) with supervised losses. Unsupervised losses measured with the kernel dependence criterion MMD enforce Gaussianity of $\boldsymbol{z}$ and independence (factorization) of the joint probability of the extended space $p(\hat{\boldsymbol{y}}, \hat{\boldsymbol{z}})$ ($\mathcal{L}_z$), and lead generated $\boldsymbol{x}$ to follow the prior $p(\boldsymbol{x})$ ($\mathcal{L}_x$).

Once trained, the model enables retrievals through the inverse pass. An observation $\boldsymbol{y}$ and a sample of $\boldsymbol{z}$ from a standard normal distribution are input to the inverse pass to generate an estimate of $\boldsymbol{x}$. This process can be iterated multiple times with re-sampled $\boldsymbol{z}$ for the same $\boldsymbol{y}$, yielding a set of $\boldsymbol{x}$ samples conforming to the posterior distribution $p(\boldsymbol{x}|\boldsymbol{y})$ [29]. This sampling-based, non-parametric retrieval posterior comprehensively characterizes the retrieval solution. The most probable point-solution can be obtained through maximum a posteriori (MAP) estimation, and the retrieval uncertainty can be quantified using computable measures of sample distribution spread such as standard deviation or credible intervals.

C. Hyperparameter selection

INNs have many model hyperparameters that influence the learning process and whose optimal values have to be determined manually. Some of the most important are the subnetwork architecture, the number of affine coupling layers, the number of latent dimensions $\boldsymbol{z}$, the kernel choice for the MMD losses, and the loss term weights. We approach the hyperparameter optimisation process by initially jointly finding plausible ranges via random search, and then tuning each hyperparameter category individually while understand-

ing their effect on the model predictions. In the following, we propose the set of hyperparameters we arrived at for our INNs. The search was conducted using a baseline regional dataset (cf. Section III), and the optimal values found applied across all models for ease of implementation.

We first select the kernel for the MMD losses. The kernel needs to discern when samples are drawn from the same distribution, and hence, its choice of lengthscale is critical—too large and even different distributions will appear similar, too small and the losses will never converge. To be robust against such biases, we use a composite kernel function composed of the sum of three inverse multi-quadratic kernels $k(\boldsymbol{x}, \boldsymbol{x}') = 1/(1 + \|(\boldsymbol{x} - \boldsymbol{x}')/h\|^2)$ with three different scale parameters $h \in \{0.03, 0.3, 1\}$ [35], which results in a peaky kernel with heavy tails able to capture features on multiple scales.

As for the model architecture details, we find that higher parameter counts in the subnetworks correlates with prediction accuracy. We settle on a configuration with only two hidden layers and 64 neurons per layer in each subnetwork to reduce model complexity while maintaining sufficient accuracy. We use a total of 4 coupling blocks in sequence for the INN, which we find sufficient to achieve the required distribution transformations. For the number of latent dimensions $\boldsymbol{z}$, we encounter a trade-off between reconstruction accuracy and convergence of $\mathcal{L}_z$. A higher number of latent dimensions allows more accurate retrievals, as more information about $\boldsymbol{x}$ can be encoded, but deteriorates the calibration of the posterior distributions as the MMD loss pushing $\boldsymbol{z}$ to be Gaussian for sampling is more difficult to optimize. We choose a mid-range value of 10 latent variables $\boldsymbol{z}$ to balance both objectives.

The loss weights are significant for the final model quality, and our search provides insight into their specific effects. Giving high weight to both $\mathcal{L}_y$ and $\mathcal{L}_r$ is necessary to obtain high accuracy in both forward predictions and retrievals. While the INN being invertible could lead one to expect that high accuracy in the forward direction would translate directly to high accuracy in the inverse, we find that strongly weighting the reconstruction loss is still important due to the randomness introduced by the latent variables $\boldsymbol{z}$. In fact, $\mathcal{L}_y$ controls the accuracy of forward predictions of $\boldsymbol{y}$, and $\mathcal{L}_z$ is only concerned with the overall distribution of $\boldsymbol{z}$. By penalizing reconstruction error, $\mathcal{L}_r$ enforces consistency in the $\boldsymbol{z}$ values output by the same input $\boldsymbol{x}$, which contributes to learning to encode valuable features. The $\mathcal{L}_z$ loss is essential for posterior calibration, as its objective constitutes the basis for the INN's probabilistic component. The $\mathcal{L}_x$ loss only has the practical purpose of easing and accelerating the convergence of the other losses. Therefore we can get away with giving it a lower weight and controlling the learning process more directly via the learning rate or optimiser. After performing Bayesian hyperparameter searches, we choose the final values for the loss weights, optimizing for both accuracy and uncertainty calibration metrics and balancing the trade-off. The weights used are $\gamma_y = 30$, $\gamma_z = 25$, $\gamma_x = 15$, and $\gamma_r = 30$.

III. Data collection and experimental setup

A. Data sources

The data used to train an INN must include the state variables input to the forward problem, which will also be retrieved by the inverse, and the corresponding observational variables. In our case, suitable training data must include at least AOD, surface reflectance, and corresponding TOA reflectance. While data with these characteristics would be available in existing AOD satellite products, they are unsuitable for making full use of the capabilities of INNs. The data are already the result of a retrieval process, and hence, the probability distribution that the INN aims to reveal collapses to a single-point estimate. Using them for training would embed in the INN the biases of the original retrieval algorithm. Instead, we opt for using RTM-simulated data directly, which encodes the physics of the forward problem. They form the foundation for all AOD retrieval algorithms and provide appropriate ground truth for training the INNs.

RTM simulations provide the atmospheric components of the forward problem, that can be combined with surface reflectance to calculate the TOA reflectance observed by satellite sensors. We follow the definition employed by the MODIS Dark Target (DT) algorithm for AOD retrieval [11, 36], and model the gas-corrected total reflectances ρ_λ^{tot} for a given surface reflectance ρ_λ^{surf} with the following equation (where subscript λ denotes the spectral wavelength):

$$\rho_\lambda^{tot} = \rho_\lambda^{atm} + \frac{F_{d,\lambda} T_\lambda \rho_\lambda^{surf}}{1 - s_\lambda \rho_\lambda^{surf}}, \tag{10}$$

where ρ_λ^{atm}, $F_{d,\lambda}$, T_λ, and s_λ are, respectively, the atmospheric path reflectance, the normalized downward flux, the total upward transmittance, and the atmospheric backscattering ratio (spherical albedo). These terms, often called atmospheric transfer functions, are spectral magnitudes that depend on the atmospheric conditions (namely the AOD) and illumination/viewing geometry (i.e., solar zenith angle θ_0, sensor zenith angle θ, and relative azimuth angle ϕ; hereafter referred to as the angles). Equation 10 is valid for a Lambertian surface, but the error introduced by this approximation is considered negligible [37] and is reduced by using a BRDF model to calculate the surface reflectance for the specific angle configuration considered [5]. In the DT problem formulation, fine- and coarse-dominated aerosol models are considered to make up the overall aerosol signal. Accordingly, the observed TOA reflectance ρ_λ^{TOA} is modelled as a weighted sum of the fine and coarse ρ_λ^{tot}, where the proportion of the fine-model signal is named the fine model weighting (FW) [36]. Note that in the broader literature, fine-mode fraction (FMF) refers to the fractional contribution of fine-mode aerosols to AOD. The DT definition instead considers the contribution of the fine-dominated model (which still comprises some coarse-size modes) to ρ_λ^{TOA}. These differences mean that the MODIS retrieved FW cannot be easily compared with other FMF products, such as from Aeronet [38].

To generate our training datasets, we use data from MODIS DT's RTM-generated LUTs [36, 39]. These LUTs contain precomputed values of the atmospheric transfer functions in (10)

in the four MODIS channels used for the retrieval (namely at 466 nm, 553 nm, 644 nm and 2130 nm), indexed by AOD at 550 nm and angles θ_0, θ, ϕ in discrete steps. There is a LUT for each of five aerosol models defined to represent the global variability of aerosol composition and optical properties over land, of which three correspond to fine-mode types (generic/neutral, urban/non-absorbing, smoke/absorbing) and one to coarse-mode (dust/spheroid). Dark Target climatology maps indicate which fine model type should be used depending on location and season, while the dust type is used for the coarse model. The LUTs can be interpolated to provide all the atmospheric parameters of (10) at any geometric configuration and AOD level.

As for the surface reflectance, it is obtained from the MODIS surface bidirectional reflectance distribution function (BRDF) product MCD43GF [40]. The BRDF is a function that encodes a surface's characteristics and models how it reflects radiation. This product provides the parameters of the Ross-Li BRDF kernel model [41] on a global 1-km grid at daily resolution. This BRDF model allows one to calculate the directional and hemispheric surface reflectance factors at any illumination and viewing geometry. Accordingly, this product lets us calculate realistic surface reflectances for any desired location. Combined with the atmospheric transfer functions from the DT LUTs, we can generate synthetic TOA reflectance observations with (10) and so obtain our training datasets.

B. Data generation

Fig. 3 shows a diagram describing the procedure used to generate our training datasets. A data point can be generated by defining AOD, θ_0, θ, ϕ, and FW values and selecting a surface pixel from the BRDF product. The AOD, FW, and viewing angle values are randomly sampled from pre-defined prior distributions, and a surface pixel is randomly selected from a specified region (details in the following paragraphs). The LUT data is interpolated at the given AOD and angles to obtain the fine- and coarse-mode atmospheric reflectances and $F_{d,\lambda}$, T_λ, and s_λ parameters. The fine- and coarse-type LUTs are chosen based on DT's seasonal aerosol-type maps. The BRDF parameters at the selected surface pixel and the angles are used to calculate the surface reflectances following the BRDF Ross-Li model. The obtained variables allow us to calculate the total reflectance following (10). The fine- and coarse-mode total reflectances are averaged and weighted by the FW to obtain the TOA reflectance that would be observed with the defined state.

We generate location-specific datasets to train location-specific INNs, allowing us to use local surface cover and aerosol climatology information to define appropriate priors. We use the location of existing Aeronet stations [42, 43] to pinpoint our areas of interest. The locations used in this study are presented in Table I and were chosen to cover a diversity of land covers and aerosol conditions. The region centered on the NASA Goddard Space Flight Center (GSFC) station presents varied urban and vegetated surfaces and common mixed fine aerosols, constituting good conditions for a general reference.

The prior distributions used to sample the necessary input values are derived from monthly observational and climatological data. The AOD distribution is assumed to be lognormal, with mean and standard deviation from the Aeronet station's Level 2.0 overall climatology tables [42]. The FW distribution is fitted to Aeronet FMF data using a beta distribution for its 0–1 domain. Angle distributions are based on one year of MODIS Terra overpasses at the station. For the surface reflectance, we take data for the 15$^{\text{th}}$ day of every month in 2009 from the MCD43GF product. We define a circular region of a 25 km radius centred on the station and randomly sample pixels within it to supply the surface BRDF kernel parameters and calculate the surface reflectance.

TABLE I
CHARACTERISTICS OF THE LOCATIONS USED AS BASIS FOR THE DATASETS DEVELOPED, IDENTIFIED BY THE NAME OF THE CENTRAL AERONET STATION. THE AEROSOL TYPES REFER TO THE DT FINE MODELS USED, AS DERIVED FROM THE CLIMATOLOGY MAPS. THE LAND SURFACE TYPES WERE ASSESSED INDEPENDENTLY.

Station name	Location	Land surface	Aerosol type
GSFC	Eastern USA	Suburban	Urban, generic
Solar_Village	Saudi Arabia	Sandy desert	Generic, urban
Alta_Floresta	Central Brazil	Amazonian crops	Generic
Beijing	China	Urban	Urban, generic
Mongu	Zambia	Savannah	Smoke, generic
Hyytiala	Finland	Boreal forest	Generic
Belsk	Poland	Croplands	Urban

C. Experimental setup

The INNs are set up with 7 observational variables $\boldsymbol{y}$ (ρ^{TOA}_{466}, ρ^{TOA}_{553}, ρ^{TOA}_{644}, ρ^{TOA}_{2130}, θ_0, θ and ϕ) and 9 state variables $\boldsymbol{x}$ (AOD, ρ^{surf}_{466}, ρ^{surf}_{553}, ρ^{surf}_{644}, ρ^{surf}_{2130}, FW, θ_0, θ and ϕ). The state variables $\boldsymbol{x}$ act as inputs to the forward problem and are the outputs to the inverse retrieval process; vice versa for the observational variables $\boldsymbol{y}$, which after training are the only ones needed as inputs to apply the AOD retrieval algorithm. We find that including the atmospheric reflectance and radiative transfer parameters explicitly is not necessary, as the AOD at 550 nm and the angles together encode the same information (as they do in the LUTs). The angles are thus included on both sides of the model: on the $\boldsymbol{x}$ side, they are necessary to define the forward process, while on the $\boldsymbol{y}$ side, they are invaluable inputs to the retrieval process since fixing them at retrieval time greatly reduces the space of possible solutions. For AOD, in practice, we use $\log(\text{AOD})$ to impose a positivity constraint reflecting AOD's physical definition. The data are pre-processed by applying feature-wise standard scaling. We train the INNs for 100 epochs with the Adam optimizer (learning rate 10^{-3}) on 60,000-point regional datasets (Section III) using Python and the FrEIA library [44].

IV. EVALUATION METHODOLOGY

A. Metrics

We use several metrics to evaluate the INNs' retrieval performance in terms of accuracy and uncertainty calibration. For accuracy, we use the root mean squared error (RMSE) and mean absolute error (MAE) to quantify the average prediction error, the mean error (ME) to measure overall bias, and the

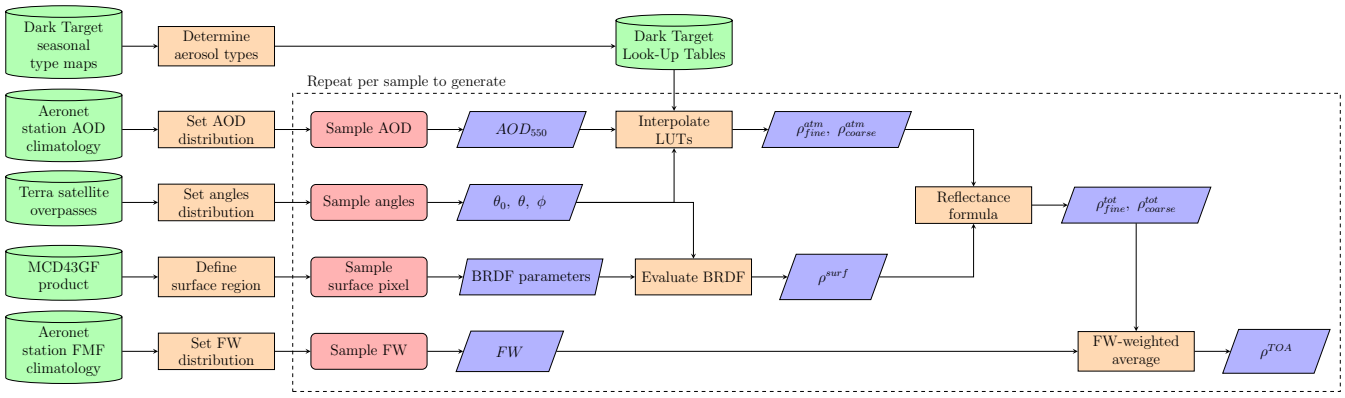

Fig. 3. Dataset generation flowchart. Green cylinders indicate data sources, orange rectangles process steps, red rounded rectangles sampling, and blue parallelograms produced variables. Monthly settings update in column two, and the dashed-line process repeats for each data point.

coefficient of determination (R^2) for a normalized measure of fit [45].

Uncertainty calibration refers to how reliable the uncertainty estimates are. For well-calibrated posteriors, a $C\%$ credible interval should contain the true value $C\%$ of the time. Over a test dataset, we denote the fraction of samples whose true value falls within the predicted credible interval C_p as C_p^{inl}, where p is the percentage covered by the interval. Then, the calibration error is defined as the difference $C_p^{inl} - C_p$ [46]. For an overall measure, we compute the mean absolute calibration error (MACE) over 100 p-confidence levels:

$$\text{MACE} = \frac{1}{100}\sum_{p=1}^{100}\left|C_p^{inl} - C_p\right|. \tag{11}$$

We are also interested in the model's predictive sharpness: for a well-calibrated model, smaller uncertainty estimates are inherently more desirable. We use the root mean squared predicted standard deviation (RMS-σ), quantifying the average spread of the posteriors:

$$\text{RMS-}\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\hat{\sigma}_i^2}, \tag{12}$$

where $\hat{\sigma}$ is the standard deviation of the predicted posterior distribution.

Finally, we report the percentage of AOD retrievals falling within the MODIS DT expected error (EE) envelope $\pm(0.05 + 15\%\text{AOD})$ [47], denoted as %EE. This is a common measure in satellite AOD product validation.

B. Synthetic test sets

We evaluate the INNs by testing their performance on synthetic test sets. We look at the GSFC region in particular for illustrative purposes. The several other regional datasets cover a diverse selection of land surface types (outlined in Table I), allowing us to investigate how surface and AOD distributions influence the retrieval. We also train and test an INN using a mixed dataset containing samples from all regions to investigate its ability to handle diverse conditions simultaneously.

Each test set consists of 24'000 points generated as previously described. For each retrieval, we generate 1000 samples for the posterior distribution and estimate the MAP solution using kernel density estimation. For the analysis, retrievals with an AOD 1-sigma uncertainty higher than 0.4 (about twice the error found for most AOD retrieval algorithms [48]) were deemed low-quality and removed. This resulted in the removal of at most 1% of data points in the Beijing set, and less than 0.1% in most others.

To establish a point of comparison with an alternative probabilistic machine learning method, we also train a fully-connected neural network (NN) to tackle the retrieval problem, using Monte Carlo (MC) dropout to estimate uncertainties [49]. We develop it for the GSFC region. The neural network with dropout (NND) uses the same $\boldsymbol{x}$ and $\boldsymbol{y}$ variables as the INN, but of course only learns the mapping from $\boldsymbol{y}$ to $\boldsymbol{x}$. We perform a random hyperparameter search optimising for accuracy and uncertainty calibration. The final NND is composed of three fully-connected hidden layers, each containing 128 neurons and a dropout probability of 0.2. It is trained using Adam optimizer with a learning rate of 0.001 for 55 epochs. The retrievals are derived with 1000 MC samples, using the mean as the point solution.

We also perform some simple experiments to explore the INNs' generalisation ability and potential to extend to real (as opposed to synthetic) conditions, and present the results in Section V-B. First, the robustness of INN retrievals to observational noise is assessed by introducing increasing levels of random Gaussian noise to the test set's TOA reflectances. The Gaussian distribution's standard deviation is derived for each measurement to achieve a certain signal-to-noise ratio (SNR) as a factor of the channel's required SNR [50]. Secondly, we explore the regional INNs' area of applicability by testing them on data generated with surfaces sampled from an increasingly larger area, which, therefore, may include examples of surface types not present in the original training set. Having trained the models with surfaces from circular regions of radius 25 km, we define larger regions with radii of 35, 50, 70, and 100 km, each roughly doubling the area.

C. Validation with real observations

The real-world accuracy of satellite AOD products is customarily validated by comparing the retrievals to collocated ground-based measurements from the Aeronet network [42]. We follow established procedures to perform a small-scale validation analysis for the INNs, applying them to full-resolution MODIS Terra L1B reflectance observations and comparing the obtained AOD retrievals to both Aeronet and the operational $10\,\mathrm{km}$ MODIS DT aerosol product [51].

Aeronet AOD is taken as ground truth for the validation of satellite retrievals. Aeronet is a global aerosol remote sensing network of sun photometers directly measuring solar radiation attenuation. Since the measurements are independent of surface reflectance, Aeronet can provide highly accurate AOD observations, with an uncertainty of about ± 0.01 in the visible range [42]. As Aeronet AOD is not directly available at $550\,\mathrm{nm}$, we interpolate it with a second-order polynomial fit in log-space from the available measurements in the $440\,\mathrm{nm}$–$870\,\mathrm{nm}$ range as in [52].

The MODIS L2 aerosol product MOD04_L2 provides AOD retrieved by the DT algorithm at nominal $10\,\mathrm{km}$ resolution. In the DT algorithm, the full-resolution $500\,\mathrm{m}$ MODIS pixels are aggregated into $10\,\mathrm{km}$ by $10\,\mathrm{km}$ (20×20) boxes. After pixel-wise cloud masking, the brightest 50% and darkest 20% of the remaining pixels by ρ^{TOA}_{2130} are discarded in a filtering step. If enough pixels remain, their spectral reflectances are averaged and used to retrieve AOD using the LUTs. The aggregation procedure is intended to increase the SNR and improve the accuracy of the AOD retrieval. As recommended, only retrievals in the highest quality assurance class are considered for the analysis.

For the INN AOD retrievals, we obtain the required ρ^{TOA}_{λ} and angles inputs from the $500\,\mathrm{m}$ resolution MODIS L1B reflectance product MOD02HKM and the geolocation product MOD03. The raw reflectances are gas corrected for water vapour, ozone and carbon dioxide absorption using the same procedure as DT [36], using gas amount data from the ERA5 reanalysis [53]. Water and cloudy pixels are masked using the MOD03 land-sea mask and MOD04_L2 cloud mask, respectively. The INNs are applied to the data at $500\,\mathrm{m}$ resolution, generating 500 samples per pixel from which the MAP point solution and uncertainty quantification are obtained.

For the validation analysis, we must find matches between the satellite and Aeronet measurements. We use the standard MODIS-Aeronet collocation procedure originally described in [54]. For each MODIS overpass, a box of 5 by 5 MOD04_L2 pixels ($50\,\mathrm{km}$ by $50\,\mathrm{km}$ in nominal size) centred on the Aeronet station is defined. The Aeronet AOD measurements taken within a temporal window of ± 30 minutes around the overpass time are considered. The match is considered valid if there are at least 5 valid DT pixels in the box and 2 Aeronet measurements in the time window. Then the MODIS spatial average and the Aeronet temporal average can be compared. For the INN AOD retrievals, we take the valid matches previously established and consider the average AOD in the corresponding 100 pixels by 100 pixels box aligned with the DT one.

We perform the analysis in our regions using all available 2009, 2010, and 2019 data to cover potentially different climatological conditions from the training sets. Note that for Mongu, Aeronet data from Mongu_Inn is used for 2019 as the station moved; for Solar Village no data is available for 2019. We evaluate the INN results against Aeronet using the accuracy and fit metrics previously described, and compare them to the corresponding DT results.

V. Results

A. Synthetic test sets

We analyse the INN performance using results from the GSFC model. First, we test the INN in the forward direction, in which the model essentially emulates the RTM and the TOA reflectance relation (10). The INN achieves almost perfect accuracy, with average ρ^{TOA}_{λ} RMSE = 0.0007 and average angles RMSE = $0.35°$, and $R^2 > 0.999$ for all variables. Such results confidently validate its ability to simulate the physical forward process.

In the inverse direction, the INN obtains the retrievals. The retrieval test set results are shown via parity plots in Fig. 4, with evaluation metrics summarised in Table II for both INN and NND. The INN retrieves AOD and surface reflectances well, obtaining R^2 values above 0.9. The AOD errors are very low (RMSE = 0.036, MAE = 0.025), with 96% of predictions falling within the EE. The NND attains similar performance to the INN in terms of accuracy (AOD RMSE = 0.037, %EE = 96%), but they differ significantly in uncertainty quantification. The INN obtains low values of MACE across all retrieved variables, indicating well-calibrated posterior distributions. These are accompanied by RMS-σ values that are generally in line with the overall errors, which suggests that the uncertainty estimates are appropriately confident. The AOD retrievals have particularly well-calibrated posteriors, with MACE = 1.5% and RMS-σ = 0.035. The uncertainty calibration can be examined in more detail with the calibration plots in Fig. 5, which plot C^{inl}_{p} against C_p over 100 p-confidence levels. If the curve falls below the 1:1 line, the uncertainty estimates are overconfident, and vice versa. For the INN predictions, the curve stays close to the 1:1 line, while the NND tends towards much more significant miscalibration. The NND predictions obtain excessively large MACE values, while also generally providing less informative posteriors as seen by their larger RMS-σ values. As seen in Fig. 5, the AOD retrievals are overconfident, with MACE = 10.5% and RMS-σ = 0.048. The quality of the INN posteriors is especially visible with the FW retrievals. Both INN and NND are unable to accurately retrieve FW (see Fig. 4). However, the INN can still provide appropriately calibrated posteriors while the NND tends to predict the same average value with a very narrow and thus overconfident uncertainty.

Looking at the error characteristics and retrieval posteriors in detail (see example in Fig. 6) can give us further insight into the INN's behaviour, which we can compare with expectations (see Section VI-A). Table III contains the correlation of the errors in retrieved AOD and ρ^{surf}_{λ}. The INN errors are strongly anti-correlated at all visible wavelengths, with

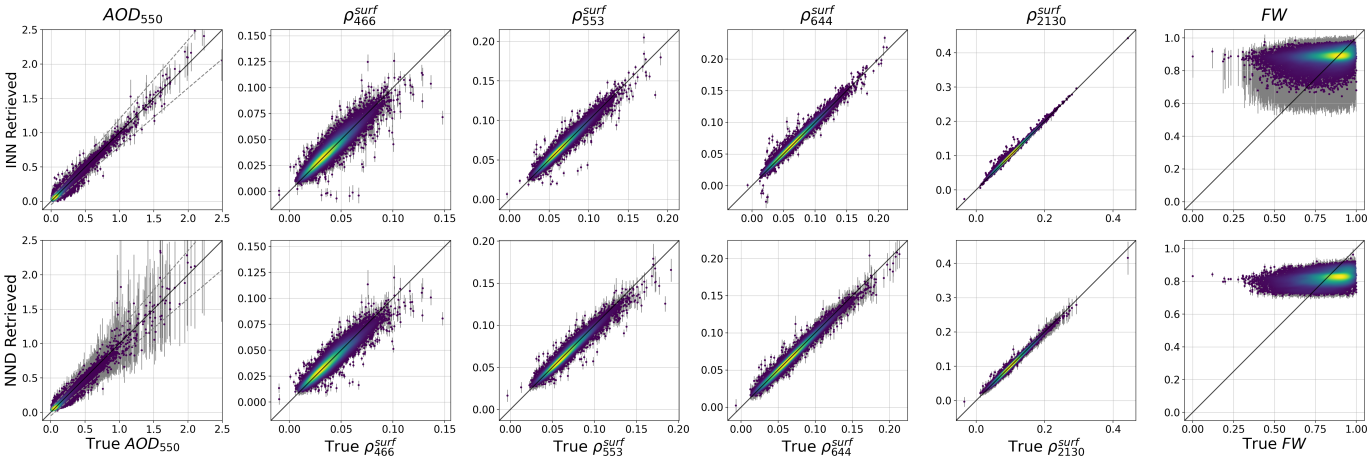

Fig. 4. Parity plots of GSFC test set retrievals with 68% uncertainty error bars, obtained with INN (top) and NND (bottom). The dashed lines in the AOD plot indicate the DT EE envelope.

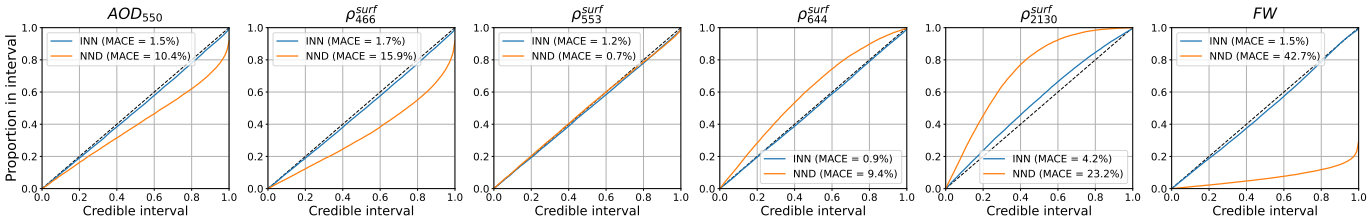

Fig. 5. Calibration plots for the GSFC test set uncertainty estimates obtained with INN and NND.

TABLE II
RETRIEVAL METRICS FOR THE GSFC TEST SET WITH INN AND NND.

	INN				NND			
Variable	RMSE	R^2	MACE	RMS-σ	RMSE	R^2	MACE	RMS-σ
AOD	0.036	0.95	1.5%	0.035	0.037	0.94	10.5%	0.048
ρ^{surf}_{λ} avg.	0.0036	0.96	2.0%	0.0030	0.0037	0.96	12.3%	0.0042
FW	0.13	-0.21	1.5%	0.11	0.11	0.12	42.7%	0.02
Angles avg.	0.40°	1.00	15.2%	0.70°	1.79°	1.00	27.5%	4.95°

a weaker relationship at $\lambda = 2130\,\mathrm{nm}$. The NND shows weaker correlations. As for the uncertainty quantification, we also show the sharpness of the ρ^{surf}_{λ} posteriors in Table III. In the INN, the uncertainty estimates show a trend of decreasing width for the longer wavelength retrievals; i.e., the greater the wavelength, the more confident the INN is in retrieving surface reflectances. Notably, the NND retrievals show the opposite trend. The posterior shapes can also reveal valuable information about the retrieval. The FW retrieval is a particularly interesting case: Fig. 6 shows that the INN's aerosol FW posterior is identical to the prior distribution given by the training set (this is reflected over the entire test set). The INN results indicate that the given input observations do not contain sufficient information to update the prior and retrieve FW. On the other hand, the NND tends to give a sharp prediction around the prior's mean value.

Table IV presents the test set metrics for every region. The accuracy results are relatively consistent throughout the

TABLE III
SURFACE REFLECTANCE RETRIEVAL POSTERIOR AND ERROR CHARACTERISTICS FOR THE GSFC TEST SET WITH INN AND NND.

	ρ^{surf}_{466}	ρ^{surf}_{553}	ρ^{surf}_{644}	ρ^{surf}_{2130}
AOD error corr.				
INN	-0.76	-0.75	-0.74	-0.37
NND	-0.59	-0.55	-0.53	-0.16
RMS-σ				
INN	0.0044	0.0032	0.0027	0.0019
NND	0.0026	0.0034	0.0042	0.0064

TABLE IV
TEST-SET AOD RETRIEVAL METRICS FOR THE DIFFERENT REGIONAL MODELS.

	Accuracy				Calibration		
Region	RMSE	MAE	ME	R^2	MACE	RMS-σ	%EE
GSFC	0.036	0.025	-0.006	0.95	1.5%	0.035	96.0%
Solar Village	0.057	0.039	-0.003	0.94	4.7%	0.047	94.3%
Alta Floresta	0.038	0.025	0.000	0.98	1.4%	0.051	96.8%
Beijing	0.084	0.055	0.002	0.97	1.7%	0.086	92.7%
Mongu	0.043	0.029	-0.008	0.95	4.2%	0.038	94.0%
Hyytiala	0.041	0.028	-0.016	0.61	2.0%	0.048	93.8%
Belsk	0.033	0.022	0.005	0.91	1.5%	0.035	96.8%
Overall	0.048	0.032	-0.004	0.90	2.4%	0.049	94.9%
Mixed regions	0.074	0.045	-0.001	0.94	3.0%	0.063	87.8%

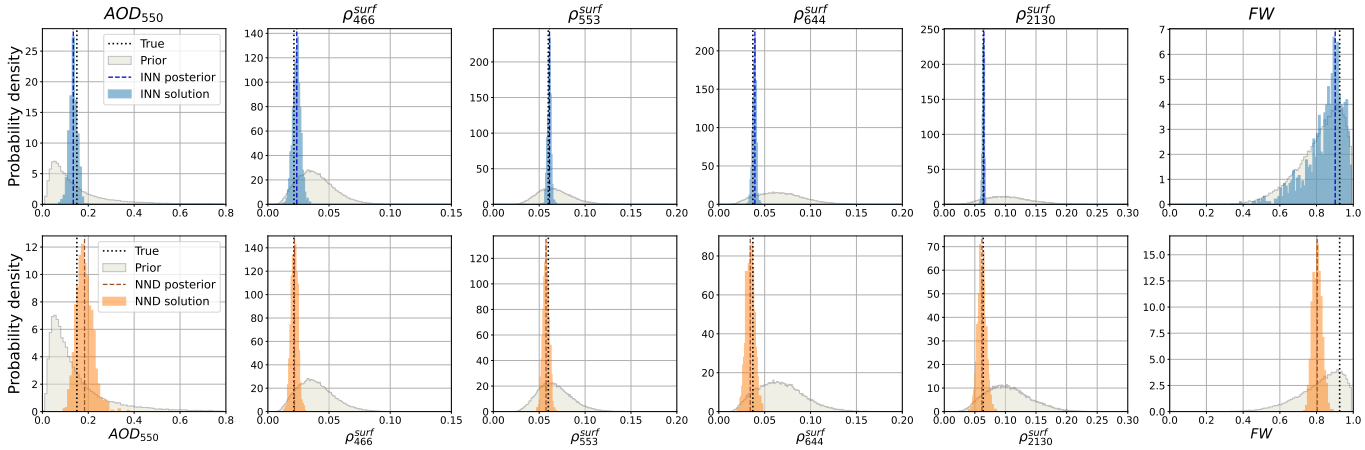

Fig. 6. Histograms of the retrieval posteriors obtained for an example observation with INN (top) and NND (bottom).

regions, with RMSE and MAE below the 0.05 threshold, except for a few notable exceptions. Beijing and Solar Village obtained worse accuracy metrics, with RMSEs of 0.084 and 0.057, respectively. Respectively a very urban environment and a dusty desert location, they are both characterised by brighter surfaces and heavier-tailed AOD distributions than the other regions. In terms of uncertainty calibration, most models achieve good MACE values below 2%. The performance is slightly worse for Solar Village (4.7%) and Mongu (4.2%). The RMS-σ values are low, mostly below 0.05, and generally align with the retrieval accuracy. Overall, the models provide both calibrated and informative uncertainty estimates.

The mixed regions model produces larger errors than the individual counterparts but still performs well overall, with %EE = 87.8%. The higher RMSE (0.074) appears to be caused mainly by larger errors in high AOD conditions, to which the MAE (0.045) is more robust. On the uncertainty estimation side, the posteriors are well calibrated (MACE = 3.0%), but this is accompanied by a slightly higher average spread (RMS-σ = 0.063), which suggests that the uncertainty estimates are slightly less informative overall.

B. Generalisation experiments

We now present the results of the two experiments examining INN generalisation ability. The results of the noise experiment for the GSFC model are depicted in Fig. 7. The curves indicate that the AOD retrieval error remains within acceptable bounds (< 0.05) even with SNR as low as a quarter of the required SNR. The calibration deteriorates earlier, but MACE remains low with SNR down to half of the required SNR.

Fig. 8 shows the change in AOD retrieval performance metrics as the radius of the region from which surfaces are sampled increases. At higher radii, the amount of points with surface characteristics not present in the original 25-km-radius training dataset is likely to increase, although the change will differ by region. As a result of the test dataset surface changes, the retrieval errors generally increase with the radius. However, the trend is not equal in all regions, and some display no

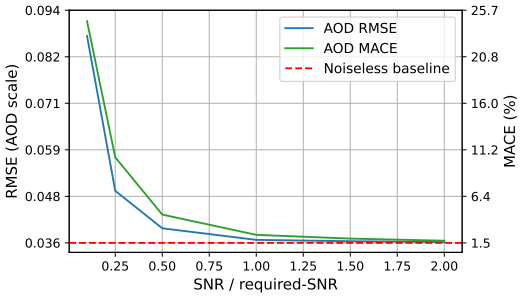

Fig. 7. AOD retrieval performance metrics in the test set as a function of required SNR noise level.

worsening of retrieval performance. Solar Village and Beijing show large RMSE increases and particularly deteriorated uncertainty calibration. MACE increases in all regions but Mongu and Hyytiala, where it remains roughly constant, as does RMSE. The difference in behaviour in each region can be related to the changes the specific ρ_λ^{surf} distributions undergo as the area is increased. In most cases, the distributions widen as new surface types are introduced (e.g. forested areas in Beijing and different urban and soil surfaces in Solar Village). Some have increasing proportions of certain surface types, eventually making the distribution multimodal (e.g. coastal waters in GSFC, deep forests in Alta Floresta). In those regions where the retrieval performance does not deteriorate, on the other hand, the ρ_λ^{surf} distributions either do not change at all, indicating a very homogeneous region (Hyytiala), or they are wider in the smallest area dataset, meaning the training set already included the full variety of encountered surface types (Mongu). It is interesting to note that the RMS-σ does not significantly change with the radius in any region. This suggests the models do not assign higher uncertainty to retrievals made from points outside their training distributions.

C. Validation with real observations

The INNs are applied to MODIS Terra L1B observations from 2009, 2010, and 2019 over the selected regions as

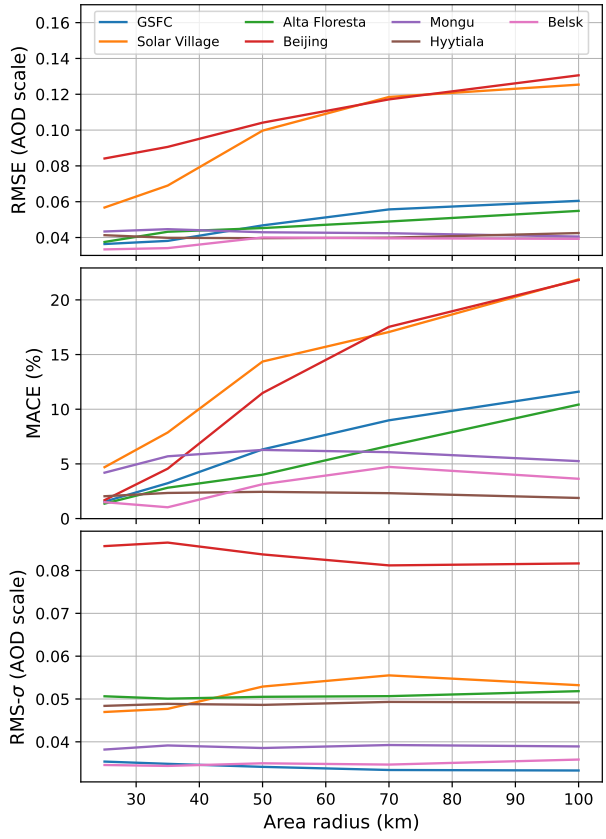

Fig. 8. AOD retrieval performance metrics in the test set as a function of the radius of the area covered for the different regions.

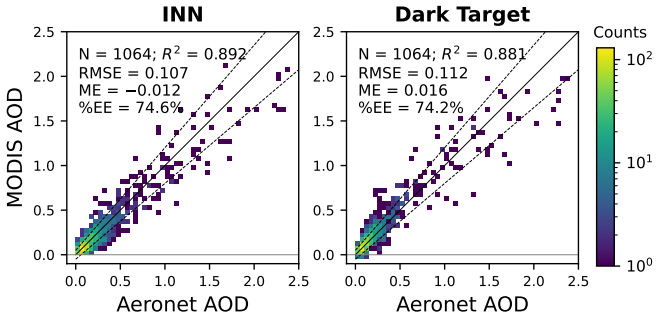

Fig. 9. Overall scatter density plots of retrieved AOD against collocated Aeronet AOD for INN and DT retrievals. The dashed lines represent the DT EE envelope.

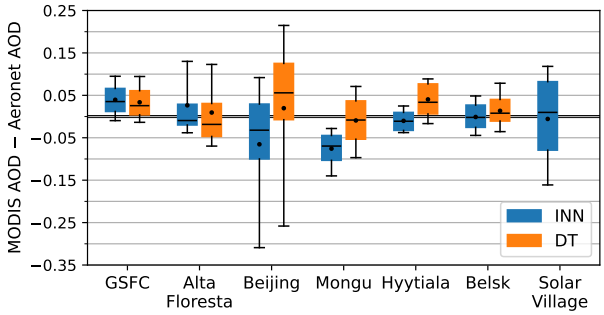

Fig. 10. Box plots summarising the distributions of AOD retrieval error by station. The box extent represents the inter-quartile range, and the whiskers the 10–90 percentile range. The midline and the dot represent the median and mean, respectively.

described in Section IV-C. The results obtained are validated against collocated Aeronet ground station measurements and the operational MODIS DT product for comparison. Key metrics are presented in Table V.

Figure 9 shows overall scatter density plots of INN and DT AOD retrievals against Aeronet. Note that it excludes data from Solar Village, as the DT algorithm does not provide retrievals in this region due to its bright sandy desert surface. Overall, INN retrievals achieve similar accuracy to DT, with RMSE $\approx$ 0.1. While the fraction of retrievals falling within the EE envelope is comparable for both algorithms (%EE $\approx$ 74%), INNs and DT display opposite bias tendencies towards under- and overestimation, respectively.

Figure 10 shows box plots summarising the distributions of INN and DT retrieval errors against Aeronet by station. The spread of the error distributions is correlated with the original width of the AOD distributions, but the bias is of comparable scale throughout. The error distributions appear tighter for the INNs than DT. The INN results are as good as DT in most regions, generally slightly better in terms of RMSE and %EE but slightly worse in ME. Exceptions are Mongu and Hyytiala, with more notably worse and better performance, respectively. INN results in Mongu display a clear negative bias, with poor RMSE and %EE. In Hyytiala, the INN demonstrates good consistency with Aeronet and outperforms DT in all metrics.

We also show validation results for the INN in the Solar Village region. DT results are not available here as the algorithm is not designed to work over bright surfaces. The INN provides skilful retrievals with low bias and good fit. RMSE and %EE are slightly worse than the average in other regions, but still reasonable and in line with DT metrics from other stations.

VI. Discussion

A. Physical interpretability

The characterising feature of INNs is that they model both the forward and inverse branches of a problem in a single model that can be run in both directions. In the forward direction, the INN essentially encodes the RTM's radiative transfer process and the TOA reflectance formula (10). This is a relatively simple deterministic problem; indeed, the INN learns to simulate it perfectly. In turn, the fact that the model now encodes the physics of the forward problem increases robustness in the inverse. In Section V-A, we presented some INN retrieval results that exhibit behaviours and relationships consistent with our physical understanding of the problem. The anti-correlation of the AOD and ρ_λ^{surf} errors reflects a physical relationship: both AOD and surface reflectance contribute positively to the TOA signal, therefore if one is underestimated the other must be overestimated and vice versa, to fit the observed ρ_λ^{TOA}. The uncertainty quantification also reflects physics. For example, the ρ_λ^{surf} posteriors have a narrower spread at longer wavelengths. Since most aerosol types have a size

TABLE V
AOD VALIDATION METRICS AGAINST AERONET BY STATION. N DENOTES THE NUMBER OF VALID COLLOCATIONS.

		RMSE		**ME**		R^2		**%EE**	
Region	N	INN	DT	INN	DT	INN	DT	INN	DT
GSFC	305	0.060	0.060	0.039	0.034	0.456	0.457	75.7%	78.7%
Solar Village	142	0.116	-	-0.006	-	0.799	-	60.6%	-
Alta Floresta	153	0.100	0.097	0.026	0.009	0.846	0.854	85.0%	73.9%
Beijing	175	0.200	0.225	-0.065	0.020	0.897	0.870	69.7%	54.3%
Mongu	220	0.096	0.075	-0.076	-0.009	0.452	0.662	53.2%	75.0%
Hyytiala	34	0.047	0.069	-0.010	0.040	0.765	0.504	88.2%	67.6%
Belsk	176	0.045	0.054	-0.001	0.013	0.786	0.693	93.2%	86.9%
Overall	1064	0.107	0.112	-0.012	0.016	0.892	0.881	74.6%	74.2%

comparable with the wavelength of visible light, scattering is strongest at shorter wavelengths and becomes negligible in the infra-red [55]. This makes ρ_λ^{surf} at longer wavelengths more easily retrievable from the TOA signal, reducing the associated uncertainty. The FW retrievals' unchanged posterior is also a sign of the INNs' reliability. Even the FW retrievals of the DT algorithm have been found to have little physical validity over land [38], and the INN can reflect this ambiguity. In these three cases, the NND demonstrates a weaker observance of physics: the error correlations are weaker, the ρ_λ^{surf} RMS-σ values follow the opposite trend, and the FW predictions are misleadingly overconfident. The results suggest that the INNs' ability to solve the inverse problem by learning to model the forward process helps enforce some physical consistency and increases reliability and trust in the predictions.

B. Out-of-distribution performance

In this study, we have focused on training INNs for AOD retrieval in limited regional contexts. As seen in Section V-B, the models can be applied beyond the original training region if the setting remains similar. Still, their performance does experience some decline when encountering surface types that fall outside the training distribution. These results help uncover a key insight into the limitations of INNs for satellite retrievals. Once trained, the INN's forward pass produces accurate results for any input since it encodes the general physical process. On the other hand, the inverse predictions and generated posteriors depend not only on the physical input observations but also on the latent variables $\boldsymbol{z}$. The latent variables learn to encode information about $\boldsymbol{x}$ during training and hence can only capture the distributions and covariances present in the training dataset. The posteriors are obtained by sampling from this latent space, which does not allow fully exploring possible solutions outside of those compatible with the learned priors. Hence, out-of-distribution observations are more likely to produce erroneous or miscalibrated retrievals.

Designing training datasets that comprehensively cover the expected conditions can enhance INNs' robustness, allowing them to perform accurately even within larger or more heterogeneous regional contexts. The idea is, in principle, also applicable in view of scaling up to a single globally applicable model. However, beyond practical challenges, following this approach could introduce some disadvantages. Since the training data distributions act as the priors for the retrievals, including more and more diverse surfaces may cause excessively complex and broad priors to be imposed on the retrievals, leading to overly broad and uninformative posteriors. This effect is hinted at in the larger RMS-σ of the mixed regions model (Table IV). A more viable approach to developing a global INN-based retrieval system would be to train different INN models for different surface types and apply them based on prior land cover knowledge. Despite such limitations in spatial scalability, INNs can be particularly useful for targeted local studies in regions of interest.

C. Application to real observations

The INNs presented in this study were developed using synthetic data generated to reflect the conditions found in a few diverse locations. The models were then applied directly to MODIS observations from those locations for validation on real data (Section V-C). The results relate to targeted contexts, but they confirm the suitability of INNs in real-world settings: INNs can retrieve AOD at full resolution with good accuracy, on the same level as the operational MODIS DT product. However, in absolute terms, the accuracy achieved is worse than suggested by the synthetic test sets; the considerations discussed in the previous section still apply. There is a distribution shift in the transition from the synthetic training data to real observations, and accounting for it can further enhance the retrieval performance in the new domain. Domain adaptation or transfer learning techniques could be applied to help close the gap between the characteristics of the synthetic and real data. For example, the INNs could be fine-tuned by further training them on MODIS DT retrievals or MODIS TOA measurements with collocated Aeronet measurements as AOD ground truth. Previous studies using post-process correction have successfully brought satellite AOD retrievals more in line with Aeronet AOD [20]. Such techniques would allow the INNs to learn the characteristics of observational data and adapt their predictions accordingly when applied to real MODIS data. They are also likely to improve the validity of the uncertainty estimates, which are more challenging to evaluate for pixel-level retrievals lacking direct ground truth. However, some approaches relying on stricter collocation with Aeronet have been proposed [52].

The availability of predictive uncertainty estimates for satellite AOD retrievals, as offered by INNs, would be an impor-

tant advantage for aerosol studies. Most traditional products do not have uncertainty quantification beyond an expected error envelope. This lack has been shown to lead to biases and increased disagreement in estimates of climate-critical aerosol-cloud interaction parameters, such as the AOD-cloud droplet number relationship, due to regression dilution [56]. Predictive uncertainty estimates can account for and mitigate this effect, improving our understanding of aerosol climate effects. Moreover, the INN-provided uncertainties are pixel-level and non-parametric, offering detailed information. They could be used to evaluate observation quality before any downstream applications. For example, the MODIS DT algorithm aggregates observations in 10 km boxes and discards 70% of pixels (assumed too bright or too dark) before proceeding with the retrieval [36]. Per-pixel uncertainties could be used to decide which pixels need to be discarded, potentially retaining more information. The INN-provided posteriors are non-parametric and hence can take any form. Their shape may carry information about the features of the aerosol being observed. For example, inspecting how the posteriors vary spatially over a scene may also help identify aerosol plumes and distinguish them from clouds, a long-standing problem for satellite retrievals.

INNs are well-suited for many satellite retrieval applications, provided the forward problem is well-defined and training data reliably links the desired parameters with radiometric observations. They would likely be even more effective when the signal in the measurements is stronger, as is the case for certain land surface parameters.

VII. Conclusion

Due to their invertible architecture and density estimation abilities, INNs are particularly suited to tackling inverse problems. We have explored their potential for probabilistic satellite aerosol retrievals, obtaining AOD with uncertainty estimates from MODIS TOA reflectance measurements. The results obtained using synthetic data demonstrate the INNs' capability to emulate and invert the atmospheric radiative transfer process, thus enabling the high-accuracy retrieval of AOD and surface reflectance. The non-parametric posterior distributions obtained for each retrieval were physically consistent in updating the priors and gave sharp and well-calibrated uncertainty estimates. Reliable posteriors can be a tool for gaining insight and trust in the model's predictions and offer added value to aerosol retrieval studies. The INNs were also applied to real MODIS observations and compared with collocated Aeronet AOD, obtaining full-resolution AOD retrievals with accuracy equivalent to that of the DT product. The INN retrievals were also effective over bright surfaces where DT is not applied.

Our analysis showed that INNs perform well when applied in the setting for which they were developed, which can make them especially useful for local studies in regions of particular interest. The training datasets define the latent space from which posteriors are generated, and they should, therefore, be carefully designed, considering the implied priors when extending the application of INNs to broader contexts. INNs trained on synthetic data can also be successfully applied to real observations, obtaining results on par with DT in their target context. Additional adjustments with domain adaptation or post-process correction techniques can further improve their performance.

In conclusion, this study demonstrated the potential of INNs as a tool for satellite aerosol retrieval, offering good accuracy with the important advantage of reliable uncertainty quantification at the pixel level. Pursuing future research directions will further enhance the utility of INNs in advancing our understanding of atmospheric aerosols and their impacts on climate.

References

[1] Y. J. Kaufman, D. Tanré, and O. Boucher, "A satellite view of aerosols in the climate system," *Nature*, vol. 419, no. 6903, pp. 215–223, Sep. 2002.

[2] IPCC, *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, United Kingdom and New York, NY, USA: Cambridge University Press, 2021.

[3] D. R. Thompson, V. Natraj, R. O. Green, M. C. Helmlinger, B.-C. Gao, and M. L. Eastwood, "Optimal estimation for imaging spectrometer atmospheric correction," *Remote Sensing of Environment*, vol. 216, pp. 355–373, 2018.

[4] Y. Kaufman, A. Wald, L. Remer, B.-C. Gao, R.-R. Li, and L. Flynn, "The MODIS 2.1-/spl mu/m channel-correlation with visible reflectance for use in remote sensing of aerosol," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 35, no. 5, pp. 1286–1298, 1997.

[5] R. C. Levy, L. A. Remer, S. Mattoo, E. F. Vermote, and Y. J. Kaufman, "Second-generation operational algorithm: Retrieval of aerosol properties over land from inversion of Moderate Resolution Imaging Spectroradiometer spectral reflectance," *Journal of Geophysical Research: Atmospheres*, vol. 112, no. D13, 2007.

[6] P. Gupta, R. C. Levy, S. Mattoo, L. A. Remer, and L. A. Munchak, "A surface reflectance scheme for retrieving aerosol optical depth over urban surfaces in MODIS Dark Target retrieval algorithm," *Atmospheric Measurement Techniques*, vol. 9, no. 7, pp. 3293–3308, 2016.

[7] N. C. Hsu, J. Lee, A. M. Sayer, W. Kim, C. Bettenhausen, and S.-C. Tsay, "VIIRS Deep Blue aerosol products over land: Extending the EOS long-term aerosol data records," *Journal of Geophysical Research: Atmospheres*, vol. 124, no. 7, pp. 4026–4053, 2019.

[8] X. Su, L. Wang, M. Zhang, W. Qin, and M. Bilal, "A high-precision aerosol retrieval algorithm (HiPARA) for Advanced Himawari Imager (AHI) data: Development and verification," *Remote Sensing of Environment*, vol. 253, p. 112221, 2021.

[9] X. Zhang, L. Li, C. Chen, X. Chen, O. Dubovik, Y. Derimian, K. Gui, Y. Zheng, H. Zhao, L. Zhang, B. Guo, Y. Wang, B. Holben, H. Che, and X. Zhang, "Validation of the aerosol optical property products derived by the GRASP/Component approach from multi-angular polarimetric observations," *Atmospheric Research*, vol. 263, p. 105802, 2021.

[10] D. Tanré, Y. J. Kaufman, M. Herman, and S. Mattoo, "Remote sensing of aerosol properties over oceans using the MODIS/EOS spectral radiances," *Journal of Geophysical Research: Atmospheres*, vol. 102, no. D14, pp. 16 971–16 988, 1997.

[11] L. A. Remer, R. C. Levy, S. Mattoo, D. Tanré, P. Gupta, Y. Shi, V. Sawyer, L. A. Munchak, Y. Zhou, M. Kim, C. Ichoku, F. Patadia, R.-R. Li, S. Gassó, R. G. Kleidman, and B. N. Holben, "The Dark Target algorithm for observing the global

aerosol system: Past, present, and future," *Remote Sensing*, vol. 12, no. 18, 2020.

[12] X. Su, L. Wang, M. Cao, L. Yang, M. Zhang, W. Qin, Q. Cao, Y. Yang, and L. Li, "Fengyun 4A land aerosol retrieval: Algorithm development, validation, and comparison with other datasets," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[13] G. E. Thomas, C. A. Poulsen, A. M. Sayer, S. H. Marsh, S. M. Dean, E. Carboni, R. Siddans, R. G. Grainger, and B. N. Lawrence, "The GRAPE aerosol retrieval algorithm," *Atmospheric Measurement Techniques*, vol. 2, no. 2, pp. 679–701, 2009.

[14] O. Dubovik, D. Fuertes, P. Litvinov, A. Lopatin, T. Lapyonok, I. Doubovik, F. Xu, F. Ducos, C. Chen, B. Torres, Y. Derimian, L. Li, M. Herreras-Giralda, M. Herrera, Y. Karol, C. Matar, G. L. Schuster, R. Espinosa, A. Puthukkudy, Z. Li, J. Fischer, R. Preusker, J. Cuesta, A. Kreuter, A. Cede, M. Aspetsberger, D. Marth, L. Bindreiter, A. Hangler, V. Lanzinger, C. Holter, and C. Federspiel, "A comprehensive description of multi-term LSM for applying multiple a priori constraints in problems of atmospheric remote sensing: GRASP algorithm, concept, and applications," *Frontiers in Remote Sensing*, vol. 2, 2021.

[15] M. Maahn, D. D. Turner, U. Löhnert, D. J. Posselt, K. Ebell, G. G. Mace, and J. M. Comstock, "Optimal estimation retrievals and their uncertainties: What every atmospheric scientist should know," *Bulletin of the American Meteorological Society*, vol. 101, no. 9, pp. E1512 – E1523, 2020.

[16] J. Vicent Servera, L. Martino, J. Verrelst, and G. Camps-Valls, "Multifidelity gaussian process emulation for atmospheric radiative transfer models," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–10, 2023.

[17] T. Su, I. Laszlo, Z. Li, J. Wei, and S. Kalluri, "Refining aerosol optical depth retrievals over land by constructing the relationship of spectral surface reflectances through deep learning: Application to Himawari-8," *Remote Sensing of Environment*, vol. 251, p. 112093, 2020.

[18] J. Lee, Y. R. Shi, C. Cai, P. Ciren, J. Wang, A. Gangopadhyay, and Z. Zhang, "Machine learning based algorithms for global dust aerosol detection from satellite images: Inter-comparisons and evaluation," *Remote Sensing*, vol. 13, no. 3, 2021.

[19] D. J. Lary, L. A. Remer, D. MacNeill, B. Roscoe, and S. Paradise, "Machine learning and bias correction of MODIS aerosol optical depth," *IEEE Geoscience and Remote Sensing Letters*, vol. 6, no. 4, pp. 694–698, 2009.

[20] A. Lipponen, J. Reinvall, A. Väisänen, H. Taskinen, T. Lähivaara, L. Sogacheva, P. Kolmonen, K. Lehtinen, A. Arola, and V. Kolehmainen, "Deep-learning-based post-process correction of the aerosol parameters in the high-resolution Sentinel-3 Level-2 Synergy product," *Atmospheric Measurement Techniques*, vol. 15, no. 4, pp. 895–914, 2022.

[21] L. A. Remer, R. C. Levy, and J. V. Martins, "Opinion: Aerosol remote sensing over the next 20 years," *Atmospheric Chemistry and Physics*, vol. 24, no. 4, pp. 2113–2127, 2024.

[22] S. Vucetic, B. Han, W. Mi, Z. Li, and Z. Obradovic, "A data-mining approach for the validation of aerosol retrievals," *IEEE Geoscience and Remote Sensing Letters*, vol. 5, no. 1, pp. 113–117, 2008.

[23] Y. Kang, M. Kim, E. Kang, D. Cho, and J. Im, "Improved retrievals of aerosol optical depth and fine mode fraction from GOCI geostationary satellite data using machine learning over East Asia," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 183, pp. 253–268, 2022.

[24] B. Han, S. Vucetic, A. Braverman, and Z. Obradovic, "A statistical complement to deterministic algorithms for the retrieval of aerosol optical thickness from radiance data," *Engineering Applications of Artificial Intelligence*, vol. 19, no. 7, pp. 787–795, 2006, special issue on Engineering Applications of Neural Networks - Novel Applications of Neural Networks in Engineering.

[25] Q. Yuan, H. Shen, T. Li, Z. Li, S. Li, Y. Jiang, H. Xu, W. Tan, Q. Yang, J. Wang, J. Gao, and L. Zhang, "Deep learning in environmental remote sensing: Achievements and challenges," *Remote Sensing of Environment*, vol. 241, p. 111716, 2020.

[26] M. Cao, M. Zhang, X. Su, and L. Wang, "A two-stage machine learning algorithm for retrieving multiple aerosol properties over land: Development and validation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–17, 2023.

[27] K. Ristovski, S. Vucetic, and Z. Obradovic, "Uncertainty analysis of neural-network-based aerosol retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 2, pp. 409–414, 2012.

[28] V. Radosavljevic, S. Vucetic, and Z. Obradovic, "A data-mining technique for aerosol retrieval across multiple accuracy measures," *IEEE Geoscience and Remote Sensing Letters*, vol. 7, no. 2, pp. 411–415, 2010.

[29] L. Ardizzone, J. Kruse, S. Wirkert, D. Rahner, E. W. Pellegrini, R. S. Klessen, L. Maier-Hein, C. Rother, and U. Köthe, "Analyzing inverse problems with invertible neural networks," 2019, arXiv:1808.04730.

[30] R. Boiger, R. L. Modini, A. Moallemi, D. Degen, A. Adelmann, and M. Gysel-Beer, "Retrieval of aerosol properties from in situ, multi-angle light scattering measurements using invertible neural networks," *Journal of Aerosol Science*, vol. 163, p. 105977, 2022.

[31] I. Kobyzev, S. J. Prince, and M. A. Brubaker, "Normalizing flows: An introduction and review of current methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, pp. 3964—3979, Nov. 2021.

[32] J. E. Johnson, V. Laparra, M. Piles, and G. Camps-Valls, "Gaussianizing the Earth: Multidimensional information measures for Earth data analysis," *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 4, pp. 191–208, 2021.

[33] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using Real NVP," 2017, arXiv:1605.08803.

[34] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, no. 25, pp. 723–773, 2012.

[35] G. Camps-Valls, L. Gómez-Chova, J. Muñoz Marí, J. Vila-Francés, and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 3, no. 1, pp. 93–97, 2006.

[36] R. C. Levy, L. A. Remer, D. Tanré, S. Mattoo, and Y. J. Kaufman, "Algorithm for remote sensing of tropospheric aerosol over dark targets for MODIS: Collections 005 and 051," MODIS Algorithm Theoretical Basis Document, 2009.

[37] Y. J. Kaufman, D. Tanré, L. A. Remer, E. F. Vermote, A. Chu, and B. N. Holben, "Operational remote sensing of tropospheric aerosol over land from EOS moderate resolution imaging spectroradiometer," *Journal of Geophysical Research: Atmospheres*, vol. 102, no. D14, pp. 17 051–17 067, 1997.

[38] R. C. Levy, L. A. Remer, R. G. Kleidman, S. Mattoo, C. Ichoku, R. Kahn, and T. F. Eck, "Global evaluation of the Collection 5 MODIS dark-target aerosol products over land," *Atmospheric Chemistry and Physics*, vol. 10, no. 21, pp. 10 399–10 420, 2010.

[39] R. Levy, "Dark Target aerosol retrieval algorithm: Stand-alone code," https://darktarget.gsfc.nasa.gov/reference/code, 2023, accessed: 2024-02-02.

[40] C. Schaaf, "MODIS/Terra+Aqua BRDF/Albedo Gap-Filled Snow-Free Daily L3 Global 30ArcSec CMG V006 [Data set]," NASA EOSDIS Land Processes Distributed Active Archive Center, 2019.

[41] W. Wanner, X. Li, and A. H. Strahler, "On the derivation of kernels for kernel-driven models of bidirectional reflectance," *Journal of Geophysical Research: Atmospheres*, vol. 100, no. D10, pp. 21 077–21 089, 1995.

[42] D. M. Giles, A. Sinyuk, M. G. Sorokin, J. S. Schafer, A. Smirnov, I. Slutsker, T. F. Eck, B. N. Holben, J. R. Lewis,

J. R. Campbell, E. J. Welton, S. V. Korkin, and A. I. Lyapustin, "Advancements in the Aerosol Robotic Network (AERONET) version 3 database – automated near-real-time quality control algorithm with improved cloud screening for sun photometer aerosol optical depth (AOD) measurements," *Atmospheric Measurement Techniques*, vol. 12, no. 1, pp. 169–209, 2019.

[43] I. Slutsker and P. Gupta, "AERONET site lists (V3)," https://aeronet.gsfc.nasa.gov/aeronet_locations_v3.txt, 2024, accessed: 2024-06-10.

[44] L. Ardizzone, T. Bungert, F. Draxler, U. Köthe, J. Kruse, R. Schmier, and P. Sorrenson, "Framework for Easily Invertible Architectures (FrEIA)," 2018-2022.

[45] G. Casella and R. L. Berger, *Statistical Inference (Second Edition)*. Duxbury/Thomson Learning, 2002.

[46] V. Kuleshov, N. Fenner, and S. Ermon, "Accurate uncertainties for deep learning using calibrated regression," 2018, arXiv:1807.00263.

[47] L. A. Remer, Y. J. Kaufman, D. Tanré, S. Mattoo, D. A. Chu, J. V. Martins, R.-R. Li, C. Ichoku, R. C. Levy, R. G. Kleidman, T. F. Eck, E. Vermote, and B. N. Holben, "The MODIS aerosol algorithm, products, and validation," *Journal of the Atmospheric Sciences*, vol. 62, no. 4, pp. 947–973, 2005.

[48] G. Doxani, E. F. Vermote, J.-C. Roger, S. Skakun, F. Gascon, A. Collison, L. De Keukelaere, C. Desjardins, D. Frantz, O. Hagolle, M. Kim, J. Louis, F. Pacifici, B. Pflug, H. Poilvé, D. Ramon, R. Richter, and F. Yin, "Atmospheric Correction Inter-comparison eXercise, ACIX-II Land: An assessment of atmospheric correction processors for Landsat 8 and Sentinel-2 over land," *Remote Sensing of Environment*, vol. 285, p. 113412, 2023.

[49] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," 2016, arXiv:1506.02142.

[50] B. Maccherone, "MODIS Technical Specifications," n.d., accessed: 2024-02-23.

[51] R. Levy, C. Hsu *et al.*, "MODIS Atmosphere L2 Aerosol Product," NASA MODIS Adaptive Processing System, Goddard Space Flight Center, USA, 2015.

[52] A. M. Sayer, Y. Govaerts, P. Kolmonen, A. Lipponen, M. Luffarelli, T. Mielonen, F. Patadia, T. Popp, A. C. Povey, K. Stebel, and M. L. Witek, "A review and framework for the evaluation of pixel-level uncertainty estimates in satellite aerosol remote sensing," *Atmospheric Measurement Techniques*, vol. 13, no. 2, pp. 373–404, 2020.

[53] H. Hersbach, B. Bell, P. Berrisford, G. Biavati, A. Horányi, J. Muñoz Sabater, J. Nicolas, C. Peubey, R. Radu, I. Rozum, D. Schepers, A. Simmons, C. Soci, D. Dee, and J.-N. Thépaut, "ERA5 hourly data on single levels from 1940 to present," Copernicus Climate Change Service (C3S) Climate Data Store (CDS), 2023, accessed: 2024-11-01.

[54] C. Ichoku, D. A. Chu, S. Mattoo, Y. J. Kaufman, L. A. Remer, D. Tanré, I. Slutsker, and B. N. Holben, "A spatio-temporal approach for global validation and analysis of MODIS aerosol products," *Geophysical Research Letters*, vol. 29, no. 12, pp. MOD1–1–MOD1–4, 2002.

[55] J. Li, B. E. Carlson, Y. L. Yung, D. Lv, J. Hansen, J. E. Penner, H. Liao, V. Ramaswamy, R. A. Kahn, P. Zhang, and et al., "Scattering and absorbing aerosols in the climate system," *Nature Reviews Earth & Environment*, vol. 3, no. 6, p. 363–379, May 2022.

[56] E. Gryspeerdt, A. C. Povey, R. G. Grainger, O. Hasekamp, N. C. Hsu, J. P. Mulcahy, A. M. Sayer, and A. Sorooshian, "Uncertainty in aerosol–cloud radiative forcing is driven by clean conditions," *Atmospheric Chemistry and Physics*, vol. 23, no. 7, pp. 4115–4122, 2023.