A GENERATIVE LIKELIHOOD FRAMEWORK FOR HIGH-RESOLUTION CLIMATE MODEL EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Next-generation high-resolution (km-scale) climate models promise unprecedented accuracy in climate projections, but realising their potential requires robust methods to quantify how well simulations align with real-world observations. Average-based metrics conventionally used for climate model evaluation ignore the physics encoded in the finescale structures of km-scale simulations. To overcome this limitation, we propose a novel, statistically principled evaluation methodology based on the likelihood function of a generative image model. Our method provides a continuous similarity metric derived from the likelihood distribution of observation and simulation snapshots, which can redefine the evaluation, intercomparison, and parameter tuning of high-resolution climate models. We demonstrate the applicability and interpretability of this method by evaluating convective clouds simulated by two state-of-the-art global km-scale models, using their outgoing infrared radiation fields. This work establishes a scalable pathway toward observation-based evaluation of next-generation climate simulations.

1 INTRODUCTION

Climate models play a crucial role in understanding and predicting the Earth's climate, providing the foundation for assessments such as the Intergovernmental Panel on Climate Change (IPCC) reports, which guide policy and societal responses to climate change (Pörtner et al., 2022). These models integrate complex interactions between the atmosphere, oceans, land, and ice to simulate how the climate responds to natural and human-induced changes (Stocker, 2011).

Global km-scale models are the frontier in climate modelling, simulating the atmosphere and ocean at unprecedented resolution, with previously inaccessible physical detail Stevens et al. (2019). They are being developed to address long-standing limitations of low-resolution models, which rely on parameterisations to approximate unresolved processes such as convection, cloud formation, and ocean eddies — approximations that drive major systematic errors and biases. By resolving these processes more explicitly, km-scale models could substantially improve the accuracy of global and regional climate projections. However, significant uncertainties remain due to parameterisations of remaining subgrid-scale processes. To isolate, understand, and reduce these biases, km-scale models need to be thoroughly evaluated.

Satellite observations are essential for evaluating km-scale models. A model that cannot reproduce the characteristics of today's climate cannot be trusted to realistically simulate future changes under increased atmospheric carbon dioxide. Km-scale climate models simulate one possible trajectory of the weather over many decades. The statistics of this simulated time series define the model's climate and should be consistent with observations. However, since weather is intrinsically stochastic, individual simulated snapshots are not expected to match observed snapshots at that exact time. Instead, the problem of climate model evaluation is to determine whether the model reproduces the statistical properties of the observed climate system.

Traditionally, climate models are assessed by comparing spatio-temporally averaged outputs to observations, using skill metrics such as mean-square error and variance (Gleckler et al., 2008; Flato et al., 2014). While informative, such metrics disregard the spatio-temporal structure of observed and simulated fields which encodes essential information about the underlying physical processes (Labe & Barnes, 2022). To improve models, performance must be explicitly linked to these physical processes, which are localised in space and time. This requires local diagnostics of the statistical

consistency between models and observations. Recent machine learning-based approaches have begun to automate the evaluation of (km-scale) climate models. However, existing studies rely on spatial or temporal averaging, or aggregate results over large regions, limiting their interpretability (e.g., Brunner & Sippel, 2023; Mooers et al., 2023).

There are strong parallels between evaluating climate simulations and assessing deep generative image models: in both cases, the goal is to determine whether the distribution of simulated data matches that of the real world. Climate modellers recognise that simulations are imperfect, and do not replicate observations exactly, but they require methods that can quantify statistical similarity. Such methods are critical for testing new parameterisation schemes, identifying which parameter choices yield realistic simulations, and comparing models that differ in modelling strategies and produce distinct outputs. Despite the growing number of km-scale models in use, the field still lacks robust and objective metrics to evaluate which models best capture the spatio-temporal structure of the climate system.

Hence, a robust evaluation metric for high-resolution climate models is needed which: (1) assesses models based on the statistics of simulated fields, without requiring paired simulations; (2) is local in time, avoiding temporal averaging; (3) is local in space, avoiding both spatial averaging or only assessing large areas at once; (4) evaluates a field directly observable (or closely related to those observable) by satellites; (5) primarily evaluates the structures present in the field, rather than trivial differences in means or other low-order statistics; and (6) provides a quantitative distance metric that enables direct comparison across different model outputs.

To address this gap, we introduce a statistically motivated evaluation metric for assessing the realism of km-scale climate models directly against snapshots of high-resolution satellite imagery. First, we reproject the observation and model datasets to a square format better suited to train generative models. Second, we train a generative model exclusively on observational data in order to learn a statistical representation of the observed climate system. Third, we compute the likelihood distribution of observational and simulated data under the trained model, and assess the realism of simulations based on the distance between the simulation and observation likelihood distributions.

We present a case study evaluation of two state-of-the-art km-scale models, the Integrated Forecasting System (IFS, Rackow et al., 2025) and the ICOsahedral Nonhydrostatic model (ICON, Hohenegger et al., 2023), against observations from NOAA's Geostationary Operational Environmental Satellite (GOES-16, Schmit & Gunshor, 2020). Our analysis focuses on convective thunderstorm clouds, which are a major source of uncertainty in climate projections (Stephens et al., 2024). Unlike traditional low-resolution models, km-scale models operate at sufficiently high resolutions to directly simulate deep convection (Stevens et al., 2019). We evaluate simulated outgoing longwave radiation (OLR), a quantity observable from satellites and commonly used as a proxy for high cloud cover and convective activity, making it well suited for investigating deep convection.

In summary, this paper makes the following contributions:

1. We introduce a dataset-agnostic, likelihood-based evaluation metric for assessing high-resolution global climate models via comparison with satellite observations.
2. We propose a general procedure for creating directly comparable observation and simulation datasets for a fine-scale focused evaluation approach, homogenising diverse spherical grid geometries and removing large-scale biases.
3. We demonstrate the utility of our method by evaluating simulated convective clouds in two high-resolution climate models, showing that our metric can (a) identify systematic biases from spatial snapshots alone, and (b) disentangle spatial and temporal sources of bias.

2 Background

Geostationary satellite observations Atmospheric observations come from diverse platforms such as surface stations, radiosondes, and aircraft, but these provide limited spatial and temporal coverage. In contrast, modern geostationary satellites deliver continuous observations over wide regions at nadir resolutions of $\leq 2\,\text{km}$ (Schmit & Gunshor, 2020; Holmlund et al., 2021), making them well suited for evaluating km-scale climate models. They measure radiances in visible and infrared bands, from which quantities such as outgoing longwave radiation, surface temperature and cloud

properties can be derived. Geostationary sensors sample the Earth on a grid that is regular in satellite viewing geometry but projects to a curvilinear latitude-longitude grid, with highest resolution near the sub-satellite point and coarser resolution toward the limb.

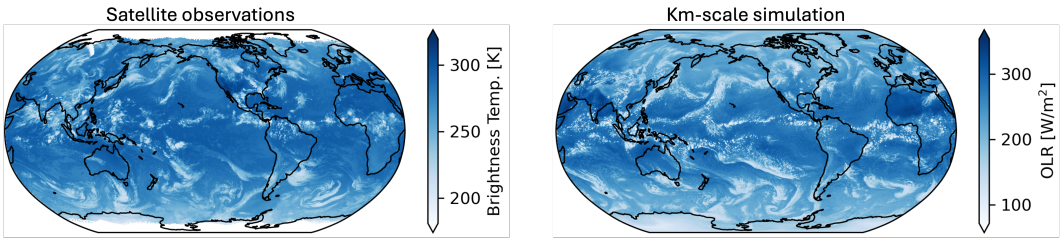

Figure 1: Example high-resolution snapshots of satellite observations and climate model simulations. Left: globally merged geostationary satellite image (11μm brightness temperature) from the preliminary ISCCP-ng dataset (CIMSS, 2025). Right: outgoing longwave radiation (OLR) field simulated by the nextGEMS ICON model (Segura et al., 2025).

Geospatial data representation A key challenge in applying machine learning to atmospheric and climate data is the representation, standardisation, and projection of input fields. While atmospheric variables are naturally defined on the spherical Earth, most machine learning frameworks operate on rectilinear arrays. Data sources add further inconsistency: climate models employ diverse non-rectilinear grids (e.g., octahedral (Rackow et al., 2025), icosahedral (Hohenegger et al., 2023)), and satellites produce instrument-specific projections (e.g., derived from the viewing geometry of a geostationary satellite). The lack of standardisation across Earth observation and climate model outputs is one of the main bottlenecks for applying machine learning methods at scale (Francis & Czerkawski, 2024). To enable meaningful comparisons between models and observations, datasets are often remapped to regular latitude–longitude grids. However, this introduces systematic distortions: pixels at higher latitudes cover smaller surface areas, inflating sampling density and biasing evaluation metrics. More suitable projections are therefore required to ensure consistent analysis.

HEALPix map projection The HEALPix (Hierarchical Equal Area isoLatitude Pixelization) scheme defines the sphere as 12 equal-area base pixels, recursively subdivided by powers of two into a quasi-regular, curvilinear grid (Górski et al., 2005). Each pixel represents an identical surface area, supporting consistent area-based metrics and fair comparisons across spatial domains. HEALPix also defines a local square coordinate structure at each subdivision level, making it compatible with standard machine learning architectures that expect rectilinear arrays. Originally developed for astronomy, HEALPix is increasingly used in atmospheric science to integrate heterogeneous observational and model datasets (Segura et al., 2025).

Normalising flows Normalising flows are a family of generative models that use a sequence of invertible and differentiable transformations to map a simple base distribution (e.g., a standard normal) into a complex target distribution matching the data of interest. Unlike other generative models such as GANs or VAEs, which provide only implicit or approximate likelihoods, normalising flows offer both tractable likelihood evaluation and efficient sampling. Early architectures such as NICE (Dinh et al., 2015) and RealNVP (Dinh et al., 2017) demonstrated flows for density estimation, while GLOW (Kingma & Dhariwal, 2018) scaled them to high-dimensional image domains. Neural Spline Flows (Durkan et al., 2019) introduced flexible, monotonic spline-based transformations that further improved expressivity while more recently, TarFlow (Zhai et al., 2025) showed that transformer-based normalising flows can achieve state-of-the-art likelihood performance and high sample quality in image generation. In this work, we adopt Neural Spline Flows, which provide a favourable balance of expressivity and parameter efficiency.

Machine learning for climate model evaluation The first applications of machine learning to climate model evaluation demonstrate the potential of data-driven approaches for evaluating and intercomparing climate models. Labe & Barnes (2022) used neural networks to identify which model produced a given annual-mean surface temperature field, and Brunner & Sippel (2023) classified

models versus observations from global daily-mean snapshots of near surface temperature. Mooers et al. (2023) compared global km-scale models using variational autoencoders which captured physically meaningful information about convection from vertical velocity fields. However, because vertical velocity cannot be directly observed, their framework cannot incorporate comparisons to satellite data. More broadly, these studies depend on global snapshots, temporally averaged fields, or variables that are not directly observable. None satisfy the requirements for a robust evaluation metric for high-resolution models outlined in Section 1, a gap which we address in this work.

3 METHODS: LIKELIHOOD-BASED EVALUATION OF KM-SCALE MODELS

The key question that climate model evaluation aims to answer is how well a model datasets represents the real climate system. To answer this question, we need to determine how *similar* the distribution of the model data is to observational data. Since the data is high dimensional, calculating the similarity between two such datasets is not straightforward. For this purpose, we propose an evaluation framework based on the likelihood function of a generative image model (Figure 2). It is trained on an observational dataset to learn its statistical distribution, and then places simulated data within this statistical distribution for comparison. Finally, the similarity between model and observations is calculated via the distance between the likelihood distributions using symmetrised KL divergence. This produces a quantitative similarity metric suitable for evaluating km-scale models.

3.1 PRELIMINARIES

A km-scale climate model, initialized at time t_0, generates a trajectory of weather states $\mathbf{x}'_1, \mathbf{x}'_2, \ldots$ whose statistics define the simulated climate. Observations provide a corresponding sequence $\mathbf{x}_1, \mathbf{x}_2, \ldots$ representing the real climate system. Because weather is intrinsically stochastic, we cannot expect $\mathbf{x}_t = \mathbf{x}'_t$ at any given time t. Instead, the task of climate model evaluation is to assess whether the statistics of the simulated climate are consistent with those of the observed system.

Formally, we assume access to some observational dataset $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, generated from an unknown data generating distribution, $\mathbf{x}_i \sim p_{\text{obs}}(\mathbf{x})$, and a simulated dataset $\mathbf{X}' = \{\mathbf{x}'_1, \ldots, \mathbf{x}'_M\}$ drawn from a km-scale model distribution $\mathbf{x}'_i \sim q_{\text{model}}(\mathbf{x})$. In general, multiple models may be considered, each creating different datasets $\mathbf{X}'_1, \mathbf{X}'_2, \ldots, \mathbf{X}'_K$. The evaluation problem is to quantify the similarity between p_{obs} and $q_{\text{model},k}$. We take $\mathbf{x}$ to be high-dimensional ($d > 500$) and assume access to a sufficiently large number of observations N to train a deep neural network.

3.2 CREATING DIRECTLY COMPARABLE OBSERVATION–SIMULATION DATASETS

Evaluating climate models against observations requires directly comparable observation and simulation datasets that represent the same variable and lie on the same grid. For neural network applications, the grid should represent sub-regions of the globe as a contiguous matrix and provide an equal-area discretisation of the sphere to ensure global statistical consistency.

Conservative remapping of geospatial data on curvilinear grids We reproject all datasets onto the HEALPix grid (Górski et al., 2005) which satisfies the above-mentioned requirements. Simple interpolation onto a different grid at the same or lower spatial resolution can introduce artifacts that bias model–observation comparisons. To mitigate this, we employ a *first-order conservative remapping* scheme, which guarantees conservation of the reprojected quantity (Jones, 1999).

Conservative remapping is the reprojection of a quantity v of a source grid onto a destination grid based on the fractional area-overlap of the source and destination grid cells. More specifically, the remapped quantity V in target grid cell j is given by:

$$V_j = \frac{1}{A_j} \sum_i A_{ij} v_i, \tag{1}$$

where A_{ij} is the intersection area between source cell i and target cell j, A_j is the area of the target cell, and v_i is the source quantity at cell i. This will automatically ensure that the integral of v over the sphere is preserved:

$$\sum_j V_j A_j = \sum_i v_i A_i. \tag{2}$$

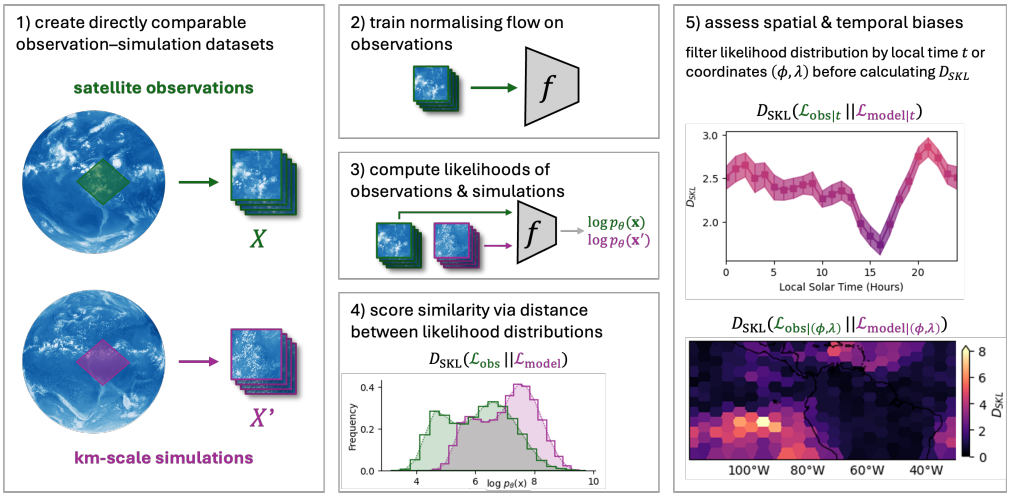

Figure 2: An overview of our likelihood-based framework for km-scale climate model evaluation. (1) We remap model and observation datasets onto the HEALPix projection to extract square patches for processing by the generative model. (2) A normalising flow model is trained on observations only and (3) used to compute the likelihood distribution of the observations and km-scale simulations. (4) We score the similarity between the simulation and observations by calculating the symmetrised KL-divergence between the likelihood distributions. (5) The likelihood distribution can be stratified by time or location to gain further insights into spatial and temporal biases.

To calculate intersection areas A_{ij}, the pixel boundaries of the source and destination grid need to be defined. For observational data such as satellite imagery, however, only the pixel (centre) coordinates are declared by the satellite's grid projection coordinates. To construct pixel boundaries, we approximate each corner as the midpoint in latitude-longitude space between the four neighbouring pixel centres on the curvilinear grid. At km-scale resolution, this approximation is accurate as pixels are sufficiently small that spherical distortions are negligible.

Removing large-scale biases via histogram matching To focus evaluation on small-scale features rather than large-scale biases, we standardise simulated data using histogram matching. Let F_{obs} and F_{sim} denote the empirical cumulative distribution functions (CDFs) of observations and simulations, respectively. Each simulated value $\mathbf{x}'$ is transformed as

$$\tilde{\mathbf{x}}' = F_{\mathrm{obs}}^{-1}(F_{\mathrm{sim}}(\mathbf{x}')) \,, \tag{3}$$

so that the transformed simulation $\tilde{\mathbf{x}}'$ follows the observed distribution. In practice, the CDFs are constructed from discretised histograms with bin width b, and the mapping is implemented by finding the smallest observation bin whose cumulative probability exceeds that of the simulated value.

3.3 Generative model likelihoods for similarity estimation

We fit a likelihood-based generative model $p(\mathbf{x};\theta)$ to the observational dataset $\mathbf{X}$, with trainable parameters θ. We use a normalizing flow, although any likelihood-based generative model could be used. The trained model provides a likelihood distribution for observational snapshots under $p(\mathbf{x};\theta)$ against which model datasets are evaluated.

Formally, we estimate discrete log likelihood distributions:

$$\mathcal{L}_{\mathrm{obs}} = \{\ell(\mathbf{x}_1), \ldots, \ell(\mathbf{x}_N)\}, \quad \ell(\mathbf{x}) = \log p(\mathbf{x};\theta), \tag{4}$$

$$\mathcal{L}_{\mathrm{model}} = \{\ell(\mathbf{x}'_1), \ldots, \ell(\mathbf{x}'_M)\}, \quad \ell(\mathbf{x}') = \log p(\mathbf{x}';\theta). \tag{5}$$

We then compute the symmetrised Kullback-Leibler (KL) divergence between $\mathcal{L}_{\text{obs}}$ and $\mathcal{L}_{\text{model}}$:

$$D_{\text{SKL}}(\mathcal{L}_{\text{obs}} \,\|\, \mathcal{L}_{\text{model}}) = \frac{1}{2}\left(D_{\text{KL}}(\mathcal{L}_{\text{obs}} \,\|\, \mathcal{L}_{\text{model}}) + D_{\text{KL}}(\mathcal{L}_{\text{model}} \,\|\, \mathcal{L}_{\text{obs}})\right) \tag{6}$$

$$= \frac{1}{2}\left(\sum_i \mathcal{L}_{\text{obs}}(i)\, \log \frac{\mathcal{L}_{\text{obs}}(i)}{\mathcal{L}_{\text{model}}(i)} + \sum_i \mathcal{L}_{\text{model}}(i)\, \log \frac{\mathcal{L}_{\text{model}}(i)}{\mathcal{L}_{\text{obs}}(i)}\right). \tag{7}$$

This divergence is zero if the two distributions are identical, and increases without bound as they diverge, thus providing a metric quantifying the similarity between observations and simulations.

Likelihoods are computed for individual patches, $\mathbf{x}$, by the generative model fitted to the whole dataset. We stratify these likelihoods by time or location to investigate temporal and spatial biases. Alongside each patch, we retain metadata: the local solar time t and the central latitude-longitude coordinates (ϕ, λ). To study temporal biases, we group by local solar time and compare the conditional likelihood distributions $\mathcal{L}_{\text{obs}|t}$ and $\mathcal{L}_{\text{model}|t}$. To study spatial biases, we group by patch centre coordinates and compare $\mathcal{L}_{\text{obs}|(\phi,\lambda)}$ and $\mathcal{L}_{\text{model}|(\phi,\lambda)}$, computing D_{SKL} within each subset.

3.4 Normalising Flow Likelihoods

Flow-based generative models define an expressive probability density on the data of interest $\mathbf{x} \in \mathbb{R}^D$ by applying an invertible, differentiable mapping $f_\theta : \mathbb{R}^D \to \mathbb{R}^D$ to a simple base random variable $\mathbf{z}$. Using the change-of-variables formula, the exact log-likelihood of a given sample $\mathbf{x}$ is:

$$\log p_\theta(\mathbf{x}) = \log p_Z(f_\theta(\mathbf{x})) + \log \left|\det \frac{\partial f_\theta(\mathbf{x})}{\partial \mathbf{x}}\right|. \tag{8}$$

4 Experiments

We use our framework for a case study evaluation of two km-scale models, IFS and ICON against observations from the geostationary satellite GOES-16. We analyse snapshots of top-of-atmosphere outgoing longwave radiation (OLR) and thereby focus our evaluation on deep convective clouds. We use PyTorch lightning for neural network training and evaluation. We extend the Neural Spline Flow implementation provided by Durkan et al. (2019) to process our OLR datasets.

4.1 Datasets and Experimental Setup

Km-scale OLR simulations We evaluate data from two global km-scale coupled models: ICON (Hohenegger et al., 2023) and IFS (Rackow et al., 2025). We analyse nextGEMS cycle 4 simulations (Segura et al., 2025), initialised with ERA5 reanalysis (Hersbach et al., 2020) at 00:00 UTC on 20 January 2020 and integrated for 30 years at $\sim$10 km atmospheric and 5 km ocean resolution. ICON directly outputs OLR as `rlut` (W/m^2). IFS provides top net thermal radiation (`ttr`) which, by definition, is equal to the negative of OLR accumulated over output intervals, i.e., over each hour (J/m^2). We thus convert `ttr` to instantaneous OLR (W/m^2) using: OLR $= -\text{ttr}/(3600 \text{ seconds})$. Both model outputs are saved on the HEALPix grid. We use the finest resolution available, HEALPix zoom level 9, with a grid spacing of $\sim 0.115^\circ \approx 12.7$ km.

GOES-16 OLR observations We evaluate simulations against observations from the GOES-16 satellite, launched in 2016 and positioned at 75.2°W. It carries the Advanced Baseline Imager (ABI) which provides full-disks image at 2 km resolution every 10 minutes (Schmit & Gunshor, 2020). We estimate OLR from ABI narrowband infrared measurements (Appendix A; Lee et al., 2010) and reproject it onto the HEALPix grid using the climate data operators conservative remapping implementation (Schulzweida, 2023).

Region, time period and train/val/test split We analyse the tropical band visible from GOES-16 which ranges from 20° to 130°W, using one year of data (2024) at hourly intervals. We split the dataset temporally into training, validation and test sets for our machine learning models. More specifically, we use days 1 to 15 of each month for training, days 20 to 23 for validation, and days 26 to 29 for testing. We leave 3 day gaps to reduce information leakage between the three datasets; this choice is motivated by the atmospheric predictability in the tropics, where small-scale ($<$100 km) features typically lose memory of their initial conditions within 5-7 days (Judt, 2020).

Table 1: Similarity scores (lower is better) based on the symmetrised KL divergence of the likelihood distribution of outgoing longwave radiation fields of two km-scale climate models IFS and ICON, compared to GOES-16 geostationary satellite observations.

	Overall	**Ocean**	**Land**
$D_{\mathrm{SKL}}(\mathcal{L}_{\mathrm{IFS}} \mid\mid \mathcal{L}_{\mathrm{GOES}})$	0.830 ± 0.013	1.650 ± 0.050	1.117 ± 0.038
$D_{\mathrm{SKL}}(\mathcal{L}_{\mathrm{ICON}} \mid\mid \mathcal{L}_{\mathrm{GOES}})$	0.148 ± 0.001	0.205 ± 0.001	0.134 ± 0.003
$D_{\mathrm{SKL}}(\mathcal{L}_{\mathrm{GOES}_{1-15}} \mid\mid \mathcal{L}_{\mathrm{GOES}_{16-31}})$	0.0004	0.001	0.003
$D_{\mathrm{SKL}}(\mathcal{L}_{\mathrm{IFS}_{1-15}} \mid\mid \mathcal{L}_{\mathrm{IFS}_{16-31}})$	0.001	0.002	0.002
$D_{\mathrm{SKL}}(\mathcal{L}_{\mathrm{ICON}_{1-15}} \mid\mid \mathcal{L}_{\mathrm{ICON}_{16-31}})$	0.001	0.002	0.002

Data Processing We empirically determine the range and distribution of values in our model and observation datasets from the training set by computing OLR histograms at a bin width of $0.5\ W/m^2$. The histograms are used to derive the cumulative distribution functions (cdfs) of our three datasets, and create lookup tables between the model and observation cdfs for histogram matching of the simulated OLR data to GOES OLR observations. Finally, we scale OLR values to the range $(0, 1)$ using the empirically determined minimum ($94.1\ W/m^2$) and maximum ($398.9\ W/m^2$) GOES OLR values. All three datasets were pre-patched to 64×64 pixel patches.

4.2 Training a Neural Spline Flow on GOES-16 Observations

We train a Neural Spline Flow model (Durkan et al., 2019) to model the GOES-16 OLR data. The architecture follows a multiscale flow with 3 levels and 7 steps per level, each step beginning with ActNorm. Transformations use rational quadratic splines with 4 bins, a tail bound of 1.0, and minimum constraints on bin width, height, and derivatives set to 10^{-3}. The coupling networks are implemented as ResNets with 3 residual blocks, 96 hidden channels, batch normalization, and no dropout. The model was trained for 20 epochs on 1 NVIDIA A100 GPU with a batch size of 64.

4.3 Quantitative evaluation of km-scale models against observations

To evaluate the realism of outgoing longwave radiation fields simulated by km-scale models, we compute the symmetrised KL divergence of the likelihood distribution between each model output and the observations. Models which replicate the observed climate distribution in the input region will have a low D_{SKL} (approaching 0) while models which fail to capture (high-resolution) features of the data distribution will have higher D_{SKL}. We additionally calculate D_{SKL} between two halves of the each dataset as a baseline for comparison. All D_{SKL} calculations in this section discretise the likelihood distributions of our observational and model datasets using 100 bins, and error bounds were estimated using bootstrap resampling.

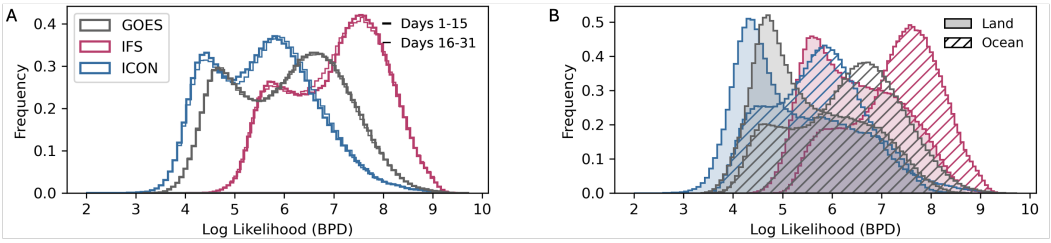

Figure 3: Histograms of log-likelihoods (bits/dim) under the neural spline flow trained on GOES satellite observations. (A) shows the likelihood distribution of GOES compared with the two km-scale simulations IFS and ICON. (B) shows likelihood distributions split into land and ocean, with patch classified as land or ocean based on its central latitude–longitude.

The likelihood distributions of both observations and simulations are bimodal (Figure 3), with the two modes corresponding to differences in cloud regimes over land and ocean. This indicates that both models capture the existence of distinct land–ocean cloud regimes, but they do not represent

each regime equally well. Both over land and over ocean, the model seems to assign particularly high likelihoods to cloud-free scenes, whereas cloudy scenes containing a lot of small-scale variability get assigned low likelihoods (Figure 8 in Appendix D.2) The two models show distinct biases (Table 1). ICON is relatively close to the observationsand shows lower divergence over land than ocean, while IFS diverges strongly. Notably, IFS scores significantly worse when the likelihood distribution is split by ocean and land. D_{SKL} between training and validation splits is very low for all datasets, confirming internal consistency.

4.4 Revealing spatial and temporal patterns of divergence

Next, we examine the spatial and temporal origins of the biases identified by our distance metric. Likelihood distributions are conditioned on patch centre coordinates to assess spatial biases, and on local solar time to assess temporal biases (Section 3.3). This stratified analysis reveals distinct spatial and temporal patterns in model errors, demonstrating the value of likelihood-based evaluation for uncovering not only overall biases but also their spatial and temporal organisation.

Figure 4 shows the divergence at each patch location. For ICON, the higher divergence over the ocean (Table 1) is concentrated in the south-western part of the domain, where deep convection is largely absent. By contrast, convectively active regions are represented exceptionally well. This indicates that ICON realistically captures deep convective structures but struggles in regimes dominated by shallow convection and clear-sky conditions. IFS, by comparison, exhibits high divergence more uniformly across the domain, with slightly larger errors in convectively active regions, pointing to systematic biases in both cloudy and clear-sky regimes.

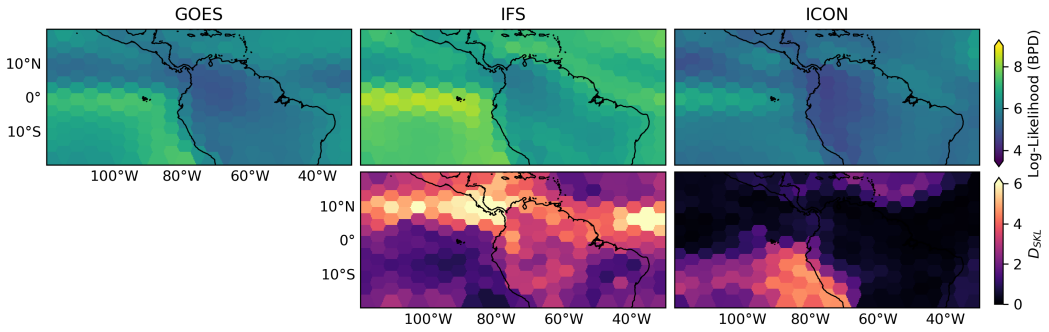

Figure 4: Analysis of spatial biases in outgoing longwave radiation of two km-scale climate models IFS and ICON, compared to GOES-16 geostationary satellite observations. Top row: maps showing the mean log-likelihood for each patch across the input region. Bottom row: maps showing the distance between likelihood distributions of IFS and ICON compared to GOES-16.

Temporal stratification exposes further structure in model biases (Figure 5). Clouds respond strongly to the diurnal cycle of incoming solar radiation, especially over land (Jones et al., 2023). Because cloud fields in turn modulate outgoing longwave radiation, their diurnal cycle is also expressed in the OLR signal. Climate models are known to struggle to capture the diurnal cycle accurately (Yin & Porporato, 2017), making its representation a critical test of model realism. Both ICON and IFS show clear time-of-day dependence in their similarity scores, with agreement generally improving in the early afternoon when convective activity peaks.

4.5 Metric Comparison

We compare our method to baseline methods and evaluate the sensitivity of our results to the size of input patches.

We calculate the OLR mean absolute error (MAE) by averaging OLR over the full year of data. In addition, we evaluate multifractal parameters of the input patches at each location, providing a cloud-sensitive measure of simulation realism that directly probes high-resolution spatial structures in the fields. Full technical details of the baseline metrics are given in Appendix D.1, with corresponding results summarized in Table 2.

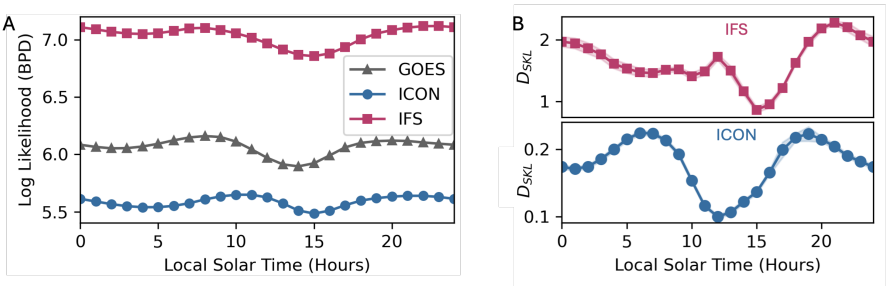

Figure 5: Analysis of temporal biases of two km-scale climate models ICON and IFS, compared to GOES-16 geostationary satellite observations. Diurnal cycle of (A) average log likelihood and (B) the distance between likelihood distributions of models and observations.

The MAE results reveal opposite biases to our likelihood-based method: for example, IFS has overall lower MAE compared to ICON, while ICON performs worse over land than over ocean. This is not unexpected, since we perform histogram matching between model and observation datasets, thereby removing mean bias to focus on small-scale structural biases. Multifractal analysis finds biases closely aligned with those discovered by our likelihood-based approach, which is encouraging given that both methods probe fine-scale variability. At the same time, our likelihood-based evaluation is more expressive, capturing model errors beyond those explained by scaling behaviour alone.

To test sensitivity to patch size, we train an additional Neural Spline Flow model on 32×32 pixel patches, using the same training setup described in Section 4.2. Results are presented in Appendix D.4. The metric scores are consistent with those obtained for 64×64 patches (Table 3). This indicates robustness to the chosen patch size and emphasises that our method can evaluate models based on small-scale structural differences.

5 Discussion

Climate model evaluation is critical for ensuring that simulations faithfully represent the Earth system and provide reliable climate projections. Traditional evaluation methods, developed for low-resolution models, rely on bias metrics or low-order statistics and therefore cannot assess the spatial and temporal structures explicitly resolved at kilometre scale.

To address this gap, we introduced a new framework that derives a quantitative similarity metric from the likelihood distribution learned by a normalising flow model. Unlike existing metrics, this approach directly measures the distance between distributions of simulated and observed snapshots. To facilitate the direct, fine-scale focused comparison between models and observations, we introduced a dataset-agnostic procedure for homogenising dataset grid projections and removing large-scale biases via histogram matching.

We present a case study evaluation of two km-scale climate models, IFS and ICON. Our results demonstrate that the likelihood-based method can robustly distinguish between models and observations, identifying spatio-temporally local biases in both of the models that were analysed. Overall, ICON exhibits closer agreement with observations across regions and the diurnal cycle. IFS has a consistent bias towards higher likelihoods, likely due to more clear-sky regions in simulated cloud fields. Thus could be due to more organised convection and thus larger structures in OLR fields which is consistent with the expected behaviour of a model that parameterises deep convection.

Our approach provides an objective, quantitative, and dataset-agnostic distance metric that captures both overall similarity and the spatial–temporal structure of model biases. This enables rigorous comparison of simulations with observations, offering guidance for the calibration of next-generation kilometre-scale climate models and help diagnose where improvements are needed. While our case study focused on outgoing longwave radiation as a proxy for cloud fields, the framework is readily extensible. Incorporating additional variables such as shortwave radiation, water vapour, or precipitation will allow a more comprehensive assessment of model realism and enable their calibration to a wide range of Earth observations.

REFERENCES

Lukas Brunner and Sebastian Sippel. Identifying climate models based on their daily output using machine learning. *Environmental Data Science*, 2:e22, 2023. doi: 10.1017/eds.2023.23.

CIMSS. International Satellite Cloud Climatology Project — Next Generation (ISCCP-NG), 2025. URL https://cimss.ssec.wisc.edu/isccp-ng/. Accessed: 2025-10-09.

Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation, 2015. URL https://arxiv.org/abs/1410.8516.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp, 2017. URL https://arxiv.org/abs/1605.08803.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows, 2019. URL https://arxiv.org/abs/1906.04032.

Robert G. Ellingson, David J. Yanuk, Hai-Tien Lee, and Arnold Gruber. A technique for estimating outgoing longwave radiation from hirs radiance observations. *Journal of Atmospheric and Oceanic Technology*, 6(4):706 – 711, 1989. doi: 10.1175/1520-0426(1989)006⟨0706:ATFEOL⟩2.0.CO;2.

Gregory Flato, Jochem Marotzke, Babatunde Abiodun, Pascale Braconnot, Sin Chan Chou, William Collins, Peter Cox, Fatima Driouech, Seita Emori, Veronika Eyring, Chris Forest, Peter Gleckler, Eric Guilyardi, Christian Jakob, Vladimir Kattsov, Chris Reason, and Markku Rummukainen. *Evaluation of Climate Models*, pp. 741–866. Cambridge University Press, 3 2014. doi: 10.1017/CBO9781107415324.020.

Alistair Francis and Mikolaj Czerkawski. Major TOM: Expandable Datasets for Earth Observation, 2024. URL https://arxiv.org/abs/2402.12095.

Lilli J. Freischem, Philipp Weiss, Hannah M. Christensen, and Philip Stier. Multifractal analysis for evaluating the representation of clouds in global kilometer-scale models. *Geophysical Research Letters*, 51(20):e2024GL110124, 2024. doi: https://doi.org/10.1029/2024GL110124. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2024GL110124.

P. J. Gleckler, K. E. Taylor, and C. Doutriaux. Performance metrics for climate models. *Journal of Geophysical Research*, 113:D06104, 3 2008. ISSN 0148-0227. doi: 10.1029/2007JD008972.

K. M. Górski, E. Hivon, A. J. Banday, B. D. Wandelt, F. K. Hansen, M. Reinecke, and M. Bartelmann. Healpix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere. *The Astrophysical Journal*, 622(2):759, apr 2005. doi: 10.1086/427976.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: https://doi.org/10.1002/qj.3803. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803.

C. Hohenegger, P. Korn, L. Linardakis, R. Redler, R. Schnur, P. Adamidis, J. Bao, S. Bastin, M. Behravesh, M. Bergemann, J. Biercamp, H. Bockelmann, R. Brokopf, N. Brüggemann, L. Casaroli, F. Chegini, G. Datseris, M. Esch, G. George, M. Giorgetta, O. Gutjahr, H. Haak, M. Hanke, T. Ilyina, T. Jahns, J. Jungclaus, M. Kern, D. Klocke, L. Kluft, T. Kölling, L. Kornblueh, S. Kosukhin, C. Kroll, J. Lee, T. Mauritsen, C. Mehlmann, T. Mieslinger, A. K. Naumann, L. Paccini, A. Peinado, D. S. Praturi, D. Putrasahan, S. Rast, T. Riddick, N. Roeber, H. Schmidt, U. Schulzweida, F. Schütte, H. Segura, R. Shevchenko, V. Singh, M. Specht, C. C. Stephan, J.-S.

von Storch, R. Vogel, C. Wengel, M. Winkler, F. Ziemen, J. Marotzke, and B. Stevens. Icon-sapphire: simulating the components of the earth system and their interactions at kilometer and subkilometer scales. *Geoscientific Model Development*, 16:779–811, 1 2023. ISSN 1991-9603. doi: 10.5194/gmd-16-779-2023.

K. Holmlund, J. Grandell, J. Schmetz, R. Stuhlmann, B. Bojkov, R. Munro, M. Lekouara, D. Coppens, B. Viticchie, T. August, B. Theodore, P. Watts, M. Dobber, G. Fowler, S. Bojinski, A. Schmid, K. Salonen, S. Tjemkes, D. Aminou, and P. Blythe. Meteosat third generation (mtg): Continuation and innovation of observations from geostationary orbit. *Bulletin of the American Meteorological Society*, 102(5):E990 – E1015, 2021. doi: 10.1175/BAMS-D-19-0304.1. URL https://journals.ametsoc.org/view/journals/bams/102/5/BAMS-D-19-0304.1.xml.

Philip W. Jones. First- and second-order conservative remapping schemes for grids in spherical coordinates. *Monthly Weather Review*, 127(9):2204 – 2210, 1999. doi: 10.1175/1520-0493(1999)127⟨2204:FASOCR⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/127/9/1520-0493_1999_127_2204_fasocr_2.0.co_2.xml.

W. K. Jones, M. Stengel, and P. Stier. A lagrangian perspective on the lifecycle and cloud radiative effect of deep convective clouds over africa. *EGUsphere*, 2023:1–25, 2023. doi: 10.5194/egusphere-2023-2059. URL https://egusphere.copernicus.org/preprints/2023/egusphere-2023-2059/.

Falko Judt. Atmospheric predictability of the tropics, middle latitudes, and polar regions explored through global storm-resolving simulations. *Journal of the Atmospheric Sciences*, 77(1):257 – 276, 2020. doi: 10.1175/JAS-D-19-0116.1. URL https://journals.ametsoc.org/view/journals/atsc/77/1/jas-d-19-0116.1.xml.

Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions, 2018. URL https://arxiv.org/abs/1807.03039.

Zachary M. Labe and Elizabeth A. Barnes. Comparison of climate model large ensembles with observations in the arctic using simple neural networks. *Earth and Space Science*, 9(7): e2022EA002348, 2022. doi: https://doi.org/10.1029/2022EA002348.

Hai-Tien Lee, Istvan Laszlo, and Arnold Gruber. ABI Earth Radiation Budget Upward Longwave Radiation: TOA (Outgoing Longwave Radiation), 2010. URL https://www.goes-r.gov/products/ATBDs/option2/RadBud_OLR_v2.0_no_color.pdf.

Griffin Mooers, Mike Pritchard, Tom Beucler, Prakhar Srivastava, Harshini Mangipudi, Liran Peng, Pierre Gentine, and Stephan Mandt. Comparing storm resolving models and climates via unsupervised machine learning. *Scientific Reports*, 13:22365, 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-49455-w.

H. O. Pörtner, D. C. Roberts, M. Tignor, E. S. Poloczanska, K. Mintenbeck, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem, and B. Rama (eds.). *IPCC, 2022: Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, UK and New York, USA, 2022. ISBN 9781009325844. doi: 10.1017/9781009325844.022.2273.

T. Rackow, X. Pedruzo-Bagazgoitia, T. Becker, S. Milinski, I. Sandu, R. Aguridan, P. Bechtold, S. Beyer, J. Bidlot, S. Boussetta, W. Deconinck, M. Diamantakis, P. Dueben, E. Dutra, R. Forbes, R. Ghosh, H. F. Goessling, I. Hadade, J. Hegewald, T. Jung, S. Keeley, L. Kluft, N. Koldunov, A. Koldunov, T. Kölling, J. Kousal, C. Kühnlein, P. Maciel, K. Mogensen, T. Quintino, I. Polichtchouk, B. Reuter, D. Sármány, P. Scholz, D. Sidorenko, J. Streffing, B. Sützl, D. Takasuka, S. Tietsche, M. Valentini, B. Vannière, N. Wedi, L. Zampieri, and F. Ziemen. Multi-year simulations at kilometre scale with the integrated forecasting system coupled to fesom2.5 and nemov3.4. *Geoscientific Model Development*, 18(1):33–69, 2025. doi: 10.5194/gmd-18-33-2025. URL https://gmd.copernicus.org/articles/18/33/2025/.

Timothy J. Schmit and Mathew M. Gunshor. Chapter 4 - ABI Imagery from the GOES-R Series. In Steven J. Goodman, Timothy J. Schmit, Jaime Daniels, and Robert J. Redmon (eds.), *The GOES-R Series*, pp. 23–34. Elsevier, 2020. ISBN 978-0-12-814327-8. doi: https://doi.org/10.1016/B978-0-12-814327-8.00004-4.

Uwe Schulzweida. CDO User Guide, October 2023. URL https://doi.org/10.5281/zenodo.10020800.

H. Segura, X. Pedruzo-Bagazgoitia, P. Weiss, S. K. Müller, T. Rackow, J. Lee, E. Dolores-Tesillos, I. Benedict, M. Aengenheyster, R. Aguridan, G. Arduini, A. J. Baker, J. Bao, S. Bastin, E. Baulenas, T. Becker, S. Beyer, H. Bockelmann, N. Brüggemann, L. Brunner, S. K. Cheedela, S. Das, J. Denissen, I. Dragaud, P. Dziekan, M. Ekblom, J. F. Engels, M. Esch, R. Forbes, C. Frauen, L. Freischem, D. García-Maroto, P. Geier, P. Gierz, Á. González-Cervera, K. Grayson, M. Griffith, O. Gutjahr, H. Haak, I. Hadade, K. Haslehner, S. ul Hasson, J. Hegewald, L. Kluft, A. Koldunov, N. Koldunov, T. Kölling, S. Koseki, S. Kosukhin, J. Kousal, P. Kuma, A. U. Kumar, R. Li, N. Maury, M. Meindl, S. Milinski, K. Mogensen, B. Niraula, J. Nowak, D. S. Praturi, U. Proske, D. Putrasahan, R. Redler, D. Santuy, D. Sármány, R. Schnur, P. Scholz, D. Sidorenko, D. Spät, B. Sützl, D. Takasuka, A. Tompkins, A. Uribe, M. Valentini, M. Veerman, A. Voigt, S. Warnau, F. Wachsmann, M. Wacławczyk, N. Wedi, K.-H. Wieners, J. Wille, M. Winkler, Y. Wu, F. Ziemen, J. Zimmermann, F. A.-M. Bender, D. Bojovic, S. Bony, S. Bordoni, P. Brehmer, M. Dengler, E. Dutra, S. Faye, E. Fischer, C. van Heerwaarden, C. Hohenegger, H. Järvinen, M. Jochum, T. Jung, J. H. Jungclaus, N. S. Keenlyside, D. Klocke, H. Konow, M. Klose, S. Malinowski, O. Martius, T. Mauritsen, J. P. Mellado, T. Mieslinger, E. Mohino, H. Pawłowska, K. Peters-von Gehlen, A. Sarré, P. Sobhani, P. Stier, L. Tuppi, P. L. Vidale, I. Sandu, and B. Stevens. nextgems: entering the era of kilometer-scale earth system modeling. *EGUsphere*, 2025:1–39, 2025. doi: 10.5194/egusphere-2025-509. URL https://egusphere.copernicus.org/preprints/2025/egusphere-2025-509/.

Graeme L. Stephens, Kathleen A. Shiro, Maria Z. Hakuba, Hanii Takahashi, Juliet A. Pilewskie, Timothy Andrews, Claudia J. Stubenrauch, and Longtao Wu. Tropical deep convection, cloud feedbacks and climate sensitivity. *Surveys in Geophysics*, 45(6):1903–1931, 2024. ISSN 1573-0956. doi: 10.1007/s10712-024-09831-1. URL https://doi.org/10.1007/s10712-024-09831-1.

Bjorn Stevens, Masaki Satoh, Ludovic Auger, Joachim Biercamp, Christopher S Bretherton, Xi Chen, Peter Düben, Falko Judt, Marat Khairoutdinov, Daniel Klocke, Chihiro Kodama, Luis Kornblueh, Shian-Jiann Lin, Philipp Neumann, William M Putman, Niklas Röber, Ryosuke Shibuya, Benoit Vanniere, Pier Luigi Vidale, Nils Wedi, and Linjiong Zhou. Dyamond: the dynamics of the atmospheric general circulation modeled on non-hydrostatic domains. *Progress in Earth and Planetary Science Stevens et al. Progress in Earth and Planetary Science*, 6:61, 2019. doi: 10.1186/s40645-019-0304-z.

Thomas Stocker. *Introduction to climate modelling*. Springer Science & Business Media, 2011.

Jun Yin and Amilcare Porporato. Diurnal cloud cycle biases in climate models. *Nature Communications*, 8(1):2269, 2017. ISSN 2041-1723. doi: 10.1038/s41467-017-02369-4. URL https://doi.org/10.1038/s41467-017-02369-4.

Shuangfei Zhai, Ruixiang Zhang, Preetum Nakkiran, David Berthelot, Jiatao Gu, Huangjie Zheng, Tianrong Chen, Miguel Angel Bautista, Navdeep Jaitly, and Josh Susskind. Normalizing flows are capable generative models, 2025. URL https://arxiv.org/abs/2412.06329.

A Appendix

A GOES-16 ABI Outgoing Longwave Radiation (OLR)

The multi-spectral outgoing longwave radiation (OLR) algorithm is based on work by Ellingson et al. (1989) and computes OLR as a weighted sum of narrowband radiances:

$$\mathrm{OLR} = a_0(\theta) + \sum_{i=1}^{n} a_i(\theta) N_i(\theta), \tag{9}$$

where a_0 is a constant regression coefficient, a_i are regression coefficients for the ith predictor, N_i is the ABI radiance of the ith predictor, and θ is the local zenith angle. Used are radiance channels 8 (6.2μm), 10 (7.3μm), 11 (8.4μm), 13 (10.3μm), and 16 (13.3μm).

The Earth Radiation Budget Team of the GOES-R Algorithm Working Group computed the regression coefficients using Clouds and the Earth's Radiant Energy System (CERES) OLR observations and OLR estimated from Spinning Enhanced Visible and Infrared Imager (SEVIRI) radiance observations (Lee et al., 2010). The SEVIRI channels used for fitting match the wavelength of the ABI channels used for GOES OLR retrievals.

B Data Details and Access Links

The km-scale climate model outputs used in our analyses have no missing data. However, GOES-16 radiance observations can have missing pixels, or be unavailable for some time steps. If one of the channels required for the OLR retrieval algorithm is missing, we cannot obtain OLR observations. In total, out of the total 8784 time steps, 123 snapshots (1.4%) could not be retrieved and were thus not considered in our analyses.

NextGEMS production simulations for ICON and IFS are archived by the German Climate Computing Center (DKRZ) and can be accessed via DKRZ's supercomputer Levante after registration at https://luv.dkrz.de/register/. GOES-16 OLR data was derived from Level 1b radiance measurements which were supplied by the National Oceanic and Atmospheric Administration (NOAA) and can be downloaded at https://console.cloud.google.com/marketplace/product/noaa-public/goes.

C HEALPIX GRID

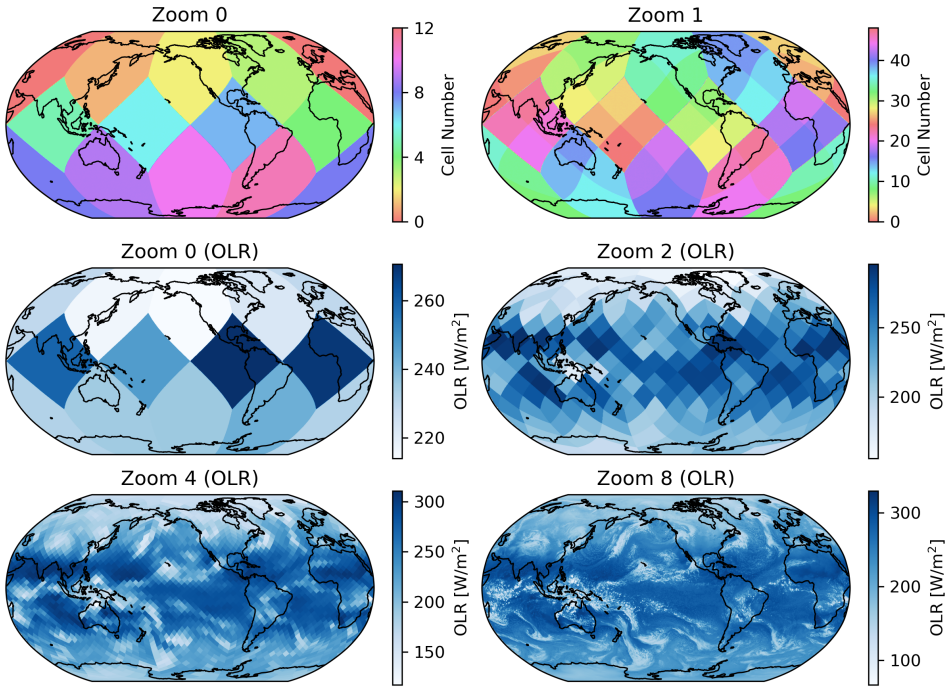

Figure 6: Visualisation of the HEALPix projection. Top row: healpix cell numbers for zoom levels 0 and 1, Second and Third row: outgoing longwave radiation (OLR) on zoom levels 1, 2, 4, 8.

D FULL LIKELIHOOD-BASED SIMILARITY METRIC RESULTS

D.1 BASELINE CLIMATE MODEL EVALUATION METRICS

Commonly used metrics for climate model evaluation include the mean absolute error (MAE) applied to long term means of the data Gleckler et al. (2008). We calculate the mean OLR for each patch location in the dataset, and compute MAE of each model as the difference between model OLR at that patch location compared to GOES OLR. This mean difference is averaged, either over the entire input region ('Overall'), or separately for patches centred over the ocean and patches centred over land.

In addition, we compare our metric to multifractal biases. Multifractal analysis is a more experimental, high-resolution focused evaluation methodology to assess the realism of simulated convective clouds in km-scale models based on their scaling behaviour (Freischem et al., 2024). We assess the error in fractal parameter ζ_∞, which is calculated as described in Freischem et al. (2024). More specifically, for each patch, we compute OLR structure functions of orders $Q = 1$ to 10 for pixel distances $r = 1$ to 40. We average structure functions across all patches at location (ϕ, λ) for the entire year, before calculating fractal parameter ζ_∞ as a fit to structure functions in range $r \in (8, 20)$. The multifractal bias at each patch location is calculated as the absolute difference in ζ_∞ between model and observations.

Table 2: Biases identified by our likelihood based metric, $D_{\mathrm{SKL}}(\mathcal{L}_{\mathrm{model}}||\mathcal{L}_{\mathrm{obs}})$, compared to mean absolute error (MAE), and fractal scaling, all evaluated on OLR fields. For D_{SKL}, MAE, and difference in fractal scaling parameters, smaller is better.

	$D_{\mathrm{SKL}}(\mathcal{L}_{\mathrm{model}} \,\|\|\, \mathcal{L}_{\mathrm{obs}}) \downarrow$			**MAE** $\downarrow$			**Multifractal** $\downarrow$		
	Overall	Ocean	Land	Overall	Ocean	Land	Overall	Ocean	Land
IFS	0.963	1.399	1.441	5.690	5.835	5.329	1.450	1.182	2.118
ICON	0.102	0.149	0.095	8.026	7.644	8.974	1.271	1.460	0.800

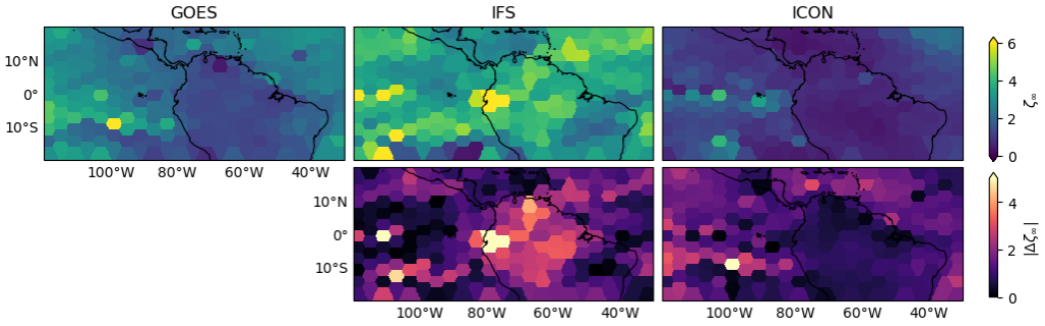

Figure 7: Multifractal parameter (top) and bias compared to GOES (bottom) on an individual patch basis.

D.2 Example Patches with Likelihoods

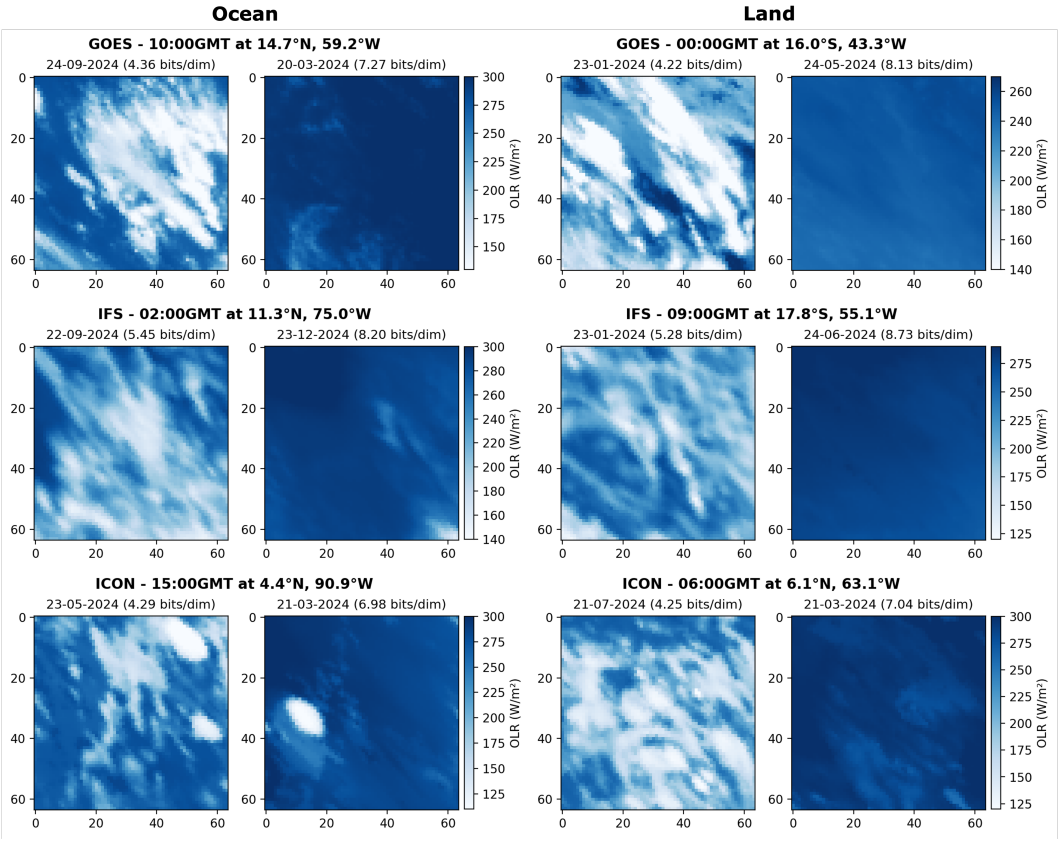

Figure 8: Example patches with low and high log likelihoods from our three datasets: (top row) GOES satellite observations, (middle row) IFS model simulations and (bottom row) ICON model simulations.

D.3 CHANGING BIASES THROUGHOUT THE DAY

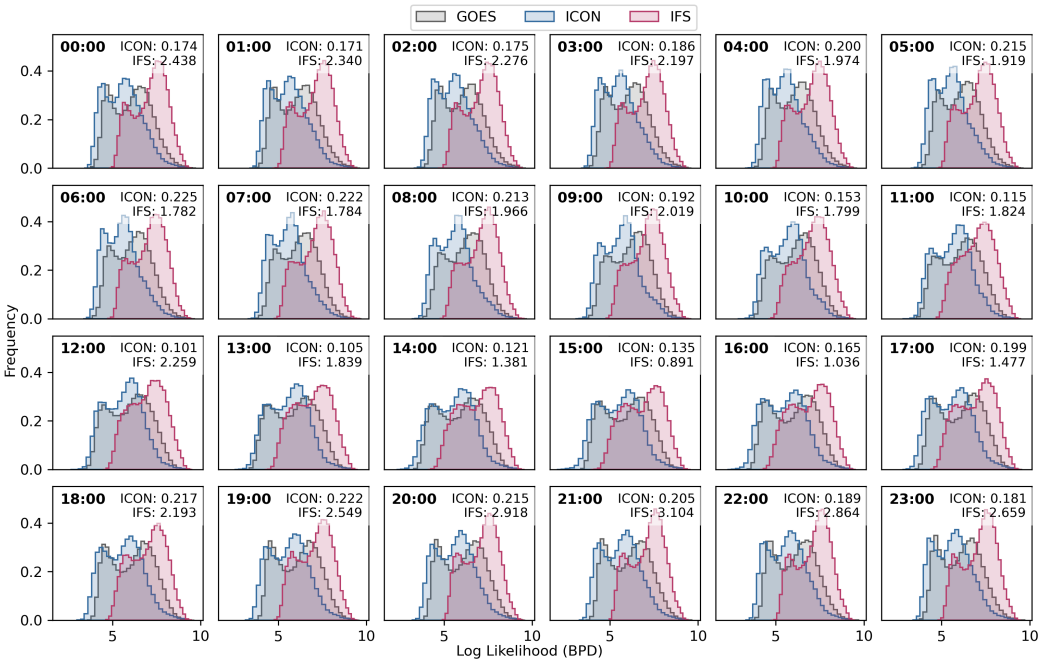

Figure 9: Log-likelihood distribution by local solar hour of the GOES-16 geostationary satellite observations, compared to the two high-resolution climate models IFS and ICON.

D.4 SENSITIVITY TEST TO PATCH SIZE

Table 3: Similarity scores (lower is better) for 32x32 pixel patches based on the symmetrised KL divergence of the likelihood distribution of outgoing longwave radiation fields of two km-scale climate models IFS and ICON, compared to GOES-16 geostationary satellite observations.

	Overall	**Ocean**	**Land**
$D_{\mathrm{SKL}}(\mathcal{L}_{\mathrm{IFS}} \,\|\|\, \mathcal{L}_{\mathrm{GOES}})$	0.774 ± 0.018	1.163 ± 0.067	1.001 ± 0.033
$D_{\mathrm{SKL}}(\mathcal{L}_{\mathrm{ICON}} \,\|\|\, \mathcal{L}_{\mathrm{GOES}})$	0.099 ± 0.001	0.161 ± 0.001	0.068 ± 0.001
$D_{\mathrm{SKL}}(\mathcal{L}_{\mathrm{GOES}_{1-15}} \,\|\|\, \mathcal{L}_{\mathrm{GOES}_{16-31}})$	0.0002	0.001	0.002
$D_{\mathrm{SKL}}(\mathcal{L}_{\mathrm{IFS}_{1-15}} \,\|\|\, \mathcal{L}_{\mathrm{IFS}_{16-31}})$	0.0009	0.001	0.002
$D_{\mathrm{SKL}}(\mathcal{L}_{\mathrm{ICON}_{1-15}} \,\|\|\, \mathcal{L}_{\mathrm{ICON}_{16-31}})$	0.0008	0.001	0.001