Environmental Data Science (2026), 5: e20, 1–12
doi:10.1017/eds.2026.10054

METHODS PAPER

A generative likelihood framework for high-resolution climate model evaluation

Lilli Johanna Freischem[1], Tim Reichelt[1], Ronald Clark[2], Philip Stier[1] and Hannah Christensen[1]

[1]Atmospheric, Oceanic and Planetary Physics, Department of Physics, University of Oxford, UK
[2]Department of Computer Science, University of Oxford, UK
Corresponding author. Lilli Johanna Freischem; Email: lilli.freischem@physics.ox.ac.uk

Received: 11 June 2026 UTC; **Revised:** 11 June 2026 UTC; **Accepted:** 14 June 2026 UTC

Keywords: climate model evaluation; deep convective clouds; km-scale climate models; normalizing flows

Abstract

Next-generation high-resolution (km-scale) climate models promise unprecedented accuracy in climate projections, but realizing 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 fine-scale 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.

Impact Statement

This new likelihood-based metric enables assessing physically relevant fine-scale structures in high-resolution climate simulations, complementing existing evaluation methods and guiding the improvement of next-generation climate models. This can enable more accurate climate projections, better suited to guide policy, and societal responses to climate change.

1. Introduction

Global km-scale models are the frontier in climate modeling, simulating the atmosphere and ocean at an unprecedented resolution of 10 km or less, with previously inaccessible physical detail (Stevens et al., 2019). They are being developed to address long-standing limitations of low-resolution models, which operate at grid spacings of around 100 km and rely on parameterizations 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

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

substantially improve the accuracy of global and regional climate projections. However, significant uncertainties remain due to parameterizations 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. Kilometer-scale models simulate one possible trajectory of the weather over many decades. The spatial and temporal statistics of this simulated time series dataset 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, variance, or correlation (Gleckler et al., 2008; Flato et al., 2013). Spectral analyses directly compare variance across spatial and temporal scales (e.g., Fraedrich and Blender, 2003; Fredriksen and Rypdal, 2016). While informative, such metrics cannot account for the inherent randomness of climate fields and disregard the high-resolution spatio-temporal structure that encodes essential information about underlying physical processes (Labe and Barnes, 2022). More comprehensive statistical approaches compare models and observations as random processes or spatial fields but are limited to one-dimensional time series, long-term averaged fields, or coarse spatial scales (e.g., Lund and Li, 2009; Zhang and Shao, 2015). Recent methods such as the spherical convolutional Wasserstein distance (SCWD) introduce spatially aware distributional comparisons, yet still cannot capture fine-scale spatial structures relevant for km-scale models (Garrett et al., 2024).

Process-oriented diagnostics have been developed to explicitly target key physical structures, particularly for deep convection. Organization indices such as I_{org} (Tompkins and Semie, 2017) quantify convective clustering and have been widely applied to both models and observations (e.g., Bony et al., 2020). Lagrangian tracking of mesoscale convective systems enables comparisons of storm frequency, size, and lifecycle statistics against satellite climatologies (Feng et al., 2021) but is typically limited to case studies of specific regions and sensitive to tracking algorithm design choices (Prein et al., 2024).

Recent machine-learning–based approaches have begun to automate the evaluation of (km-scale) climate models, demonstrating the potential of computer vision and representation learning methods (Labe and Barnes, 2022; Brunner and Sippel, 2023; Mooers et al., 2023). However, existing studies rely on global snapshots, spatial or temporal averaging, or variables that are not directly observable, limiting their interpretability and applicability for model evaluation against observations. To improve models, performance must be explicitly linked to physical processes that are localized in space and time, requiring local diagnostics of the statistical consistency between models and observations.

Hence, a robust evaluation metric for high-resolution climate models is needed that (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 difference metric that enables direct comparison across different model outputs.

To address this gap, we introduce a statistically motivated metric for evaluating km-scale climate models directly against snapshots of satellite imagery. We first reproject observations and model outputs onto a grid suitable to train neural networks. Then, we train a generative model on observational data to learn a statistical representation of the climate system. We evaluate km-scale simulations based on the divergence of likelihood distributions of simulations and observations, estimated from the trained model. We present a case study evaluation of two state-of-the-art km-scale models, the Integrated Forecasting System coupled to the FESOM ocean–sea ice model (IFS-FESOM, Rackow et al., 2025), and the

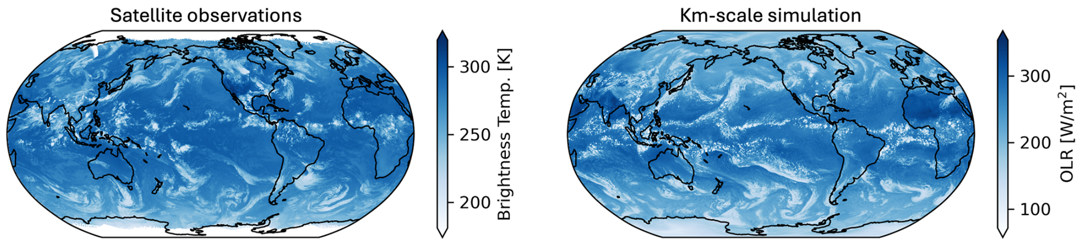

Figure 1. *Example high-resolution snapshots of satellite observations and climate model simulations. Left: globally merged geostationary satellite image (11 µm brightness temperature) from the preliminary ISCCP-ng dataset (CIMSS, 2025). Right: outgoing longwave radiation (OLR) field simulated by the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

ICOsahedral Nonhydrostatic model (ICON, Hohenegger et al., 2023), against observations from NOAA's Geostationary Operational Environmental Satellite (GOES-16, Schmit and Gunshor, 2020). We focus on convective thunderstorm clouds, which are a major source of uncertainty in climate projections (Stephens et al., 2024). We analyze outgoing longwave radiation (OLR), a quantity observable from satellites and commonly used as a proxy for high cloud cover and convective activity (Figure 1).

2. Methods: likelihood-based evaluation of km-scale models

The key question that climate model evaluation aims to answer is how well model datasets represent 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 models and observations is calculated using symmetrized Kullback–Leibler (KL) divergence of their likelihood distributions. This produces a quantitative similarity metric suitable for evaluating km-scale models.

2.1. 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. Weather is intrinsically stochastic, so 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 $\mathbf{X}' = \left\{\mathbf{x}'_1, \ldots, \mathbf{x}'_M\right\}$ drawn from a km-scale model distribution $\mathbf{x}'_i \sim q_{\text{model}}(\mathbf{x})$. Here, each sample $\mathbf{x}_i$ (and $\mathbf{x}'_i$) corresponds to a local patch rather than a global snapshot. 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 $q_{\text{model},k}$.

2.2. 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 discretization of the sphere to ensure global statistical consistency.

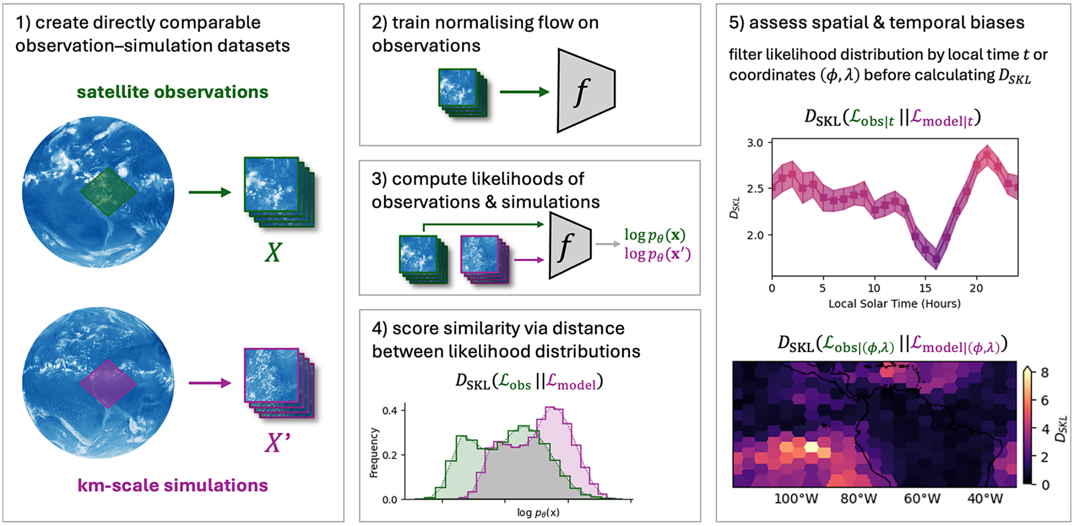

Figure 2. *An overview of our likelihood-based framework for km-scale climate model evaluation. (1) We remap model and observation datasets onto the HEALPix projection to extract square patches for processing by the generative model. (2) A normalizing flow model is trained on observations only and (3)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

2.2.1. 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. To minimize interpolation artefacts that could bias model–observation comparisons, we use a *first-order conservative remapping* scheme (Jones, 1999). This requires constructing pixel boundaries for the satellite dataset from pixel (center) coordinates. We approximate each corner as the midpoint in latitude–longitude space between the four neighboring pixel centers on the curvilinear grid. At km-scale resolution, this approximation is sufficiently accurate as spherical distortions are negligible.

2.2.2. Removing large-scale biases via histogram matching

To evaluate small-scale features rather than large-scale biases, we standardize simulated data using histogram matching. Let F_{obs} and F_{sim} denote the empirical cumulative distribution functions (CDFs) of observations and simulations, respectively, computed jointly across all spatial points. Each simulated value $\mathbf{x}'$ is transformed as $\tilde{\mathbf{x}}' = F_{\mathrm{obs}}^{-1}(F_{\mathrm{sim}}(\mathbf{x}'))$, so that the transformed simulation $\tilde{\mathbf{x}}'$ follows the observed distribution. In practice, the CDFs are constructed from discretized histograms, and the mapping is implemented by finding the smallest observation bin whose cumulative probability exceeds that of the simulated value.

2.3. 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 (see Section 2.4), although any likelihood-based generative model could be used. Here, each $\mathbf{x}_i \in \mathbf{X}$ represents a square spatial patch of the input data on the HEALPix grid. 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:

$$\mathcal{L}_{\text{obs}} = \{\log p(\mathbf{x}_1;\theta),\ldots,\log p(\mathbf{x}_N;\theta)\}, \text{ and } \mathcal{L}_{\text{model}} = \left\{\log\ p\left(\mathbf{x}'_1;\theta\right),\ldots,\log\ p\left(\mathbf{x}'_M;\theta\right)\right\}. \quad (1)$$

We then compute the symmetrized KL divergence between the two distributions of log likelihoods of the observed data, $\mathcal{L}_{\text{obs}}$, and the model data, $\mathcal{L}_{\text{model}}$:

$$D_{\text{SKL}}(\mathcal{L}_{\text{obs}} \,\|\, \mathcal{L}_{\text{model}}) = 1/2[D_{\text{KL}}(\mathcal{L}_{\text{obs}} \,\|\, \mathcal{L}_{\text{model}}) + D_{\text{KL}}(\mathcal{L}_{\text{model}} \,\|\, \mathcal{L}_{\text{obs}})] \quad (2)$$

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, which we can stratify by time or location to investigate temporal and spatial biases. Alongside each patch, we retain metadata: local solar time t and 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 coordinates and compare $\mathcal{L}_{\text{obs}|(\phi,\lambda)}$ and $\mathcal{L}_{\text{model}|(\phi,\lambda)}$, computing D_{SKL} within each subset.

2.4. Normalizing flow likelihoods

Normalizing 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 (Dinh et al., 2015; Kingma and Dhariwal, 2018). Unlike other generative models such as generative adversarial networks or variational autoencoders, which provide only implicit or approximate likelihoods, normalizing flows offer both tractable likelihood evaluation and efficient sampling.

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 $f(\cdot;\theta) : \mathbb{R}^D \to \mathbb{R}^D$ 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:

$$\log p(\mathbf{x};\theta) = \log p_Z(f(\mathbf{x};\theta)) + \log|\det(\partial f(\mathbf{x};\theta)/\partial \mathbf{x})|. \quad (3)$$

Given a training dataset $\mathcal{D} = \left\{x^{(n)}\right\}_{n=1}^{N}$ normalizing flows are trained to maximize the total log likelihood $\sum_n \log p\left(x^{(n)};\theta\right)$ with respect to θ. We use a neural spline flow (NSF) (Durkan et al., 2019), which employs monotonic rational quadratic splines to construct $f(\cdot;\theta)$ (see Section 3.1 and Supplementary Appendix C).

3. Experiments and results

We use our framework for a case study evaluation of two km-scale models, IFS-FESOM and ICON against geostationary satellite observations from GOES-16. We analyze snapshots of top-of-atmosphere OLR and thereby focus our evaluation on deep convective clouds.

We use PyTorch lightning for neural network training and evaluation. We extend the NSF implementation provided by Durkan et al. (2019) to process our OLR datasets. We compare the results of our evaluation method to three baseline metrics: mean absolute error (MAE), multifractal parameters (Freischem et al., 2024), and spherical convolutional Wasserstein distance (SCWD, Garrett et al., 2024).

3.1. Datasets and experimental setup

3.1.1. Kilometer-scale OLR simulations

We evaluate data from two global km-scale coupled models: ICON (Hohenegger et al., 2023) and IFS-FESOM (Rackow et al., 2025). Both models parametrize radiation, turbulence, and cloud microphysics, but IFS includes more parametrizations such as deep and shallow convection. ICON directly outputs OLR (W/m^2) while IFS provides top net thermal radiation (ttr) equal to negative OLR accumulated over each hour (J/m^2), which we convert to OLR using: $\text{OLR} = -\text{ttr}/(3600\,\text{s})$. We analyze nextGEMS cycle 4 simulations (Segura et al., 2025), initialized 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. 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.

3.1.2. GOES-16 OLR observations

We use 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-disk images at 2 km resolution every 10 minutes (Schmit and Gunshor, 2020). We estimate OLR from ABI narrowband infrared measurements (Supplementary Appendix A; Lee et al., 2010) and remap it onto the HEALPix grid using the climate data operators' conservative remapping tool (Schulzweida, 2023).

3.1.3. Region, time period, and train/val/test split

We analyze the tropical band visible from GOES-16, which ranges from 20° to 130°W, using 1 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 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).

3.1.4. 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 $0.5\ \mathrm{W/m^2}$. The histograms are used to derive the 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, OLR values are scaled to the range (0, 1) using empirically determined GOES minimum ($94.1\ \mathrm{W/m^2}$) and maximum ($398.9\ \mathrm{W/m^2}$) OLR values. All three datasets are patched into 64×64 pixel patches with a 32-pixel stride. For GOES, this yields 1,127,251 training and 380,781 validation patches.

3.1.5. Training a NSF on GOES-16 observations

We train an NSF (Durkan et al., 2019) to model the GOES-16 OLR data. The model was trained for 20 epochs on 1 NVIDIA A100 GPU with a batch size of 64. Trained in an unsupervised fashion to maximize dataset likelihood (Section 2.4), the model achieves average log likelihoods of 3.56 and 3.53 per pixel on the training and validation sets, respectively. Likelihoods are normalized per pixel to enable comparison across resolutions. The model assigns highest likelihoods to cloud-free patches and lowest likelihoods to patches containing popcorn convection; it generates realistic OLR patches, and sensitivity tests confirm robustness to NSF hyperparameters (Supplementary Appendix C).

3.2. Quantitative evaluation of km-scale models against observations

We evaluate the realism of OLR fields simulated by km-scale models by computing the symmetrized KL divergence of the likelihood distribution between each model output and the observations. Models that 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 each dataset for comparison. All D_{SKL} calculations in this section discretize likelihood distributions using 100 bins, and error bounds were estimated using bootstrap resampling.

The likelihood distributions of both observations and simulations are bimodal (Figure 3). The two modes correspond to clear and cloudy scenes, with cloudy being defined as patches containing over 40% cloudy pixels based on a threshold of $200\ \mathrm{W/m^2}$. Although the peaks of both modes are of similar sizes in GOES and ICON, the IFS-FESOM simulations seem to contain more clear sky than cloudy patches. The NSF assigns particularly high likelihoods to cloud-free scenes, whereas cloudy scenes with a lot of small-

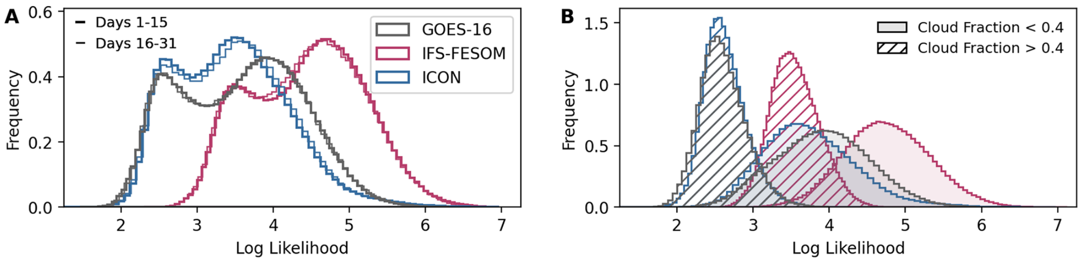

Figure 3. *Histograms of log likelihoods under the neural spline flow trained on GOES satellite data. (A) shows the likelihood distribution of GOES compared with two km-scale simulations IFS-FESOM and ICON. (B) shows likelihood distributions split by fraction of cloudy pixels per patch.*

Table 1. *Symmetrized KL divergence of likelihood distributions of outgoing longwave radiation fields of two km-scale models IFS-FESOM and ICON, compared to GOES-16 geostationary satellite observations*

	Overall	Ocean	Land
$D_{\mathrm{SKL}}(\mathcal{L}_{\mathrm{IFS-FESOM}} \| \mathcal{L}_{\mathrm{GOES}})$	1.494 ± 0.036	2.433 ± 0.044	2.332 ± 0.069
$D_{\mathrm{SKL}}(\mathcal{L}_{\mathrm{ICON}} \| \mathcal{L}_{\mathrm{GOES}})$	0.057 ± 0.001	0.110 ± 0.001	0.008 ± 0.001

scale variability get assigned the lowest likelihoods (Supplementary Appendix Figure D3). The two models show distinct biases (Table 1). ICON is closer to observations and shows lower divergence over land than ocean, while IFS-FESOM diverges strongly. Notably, IFS-FESOM scores significantly worsen 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 (Supplementary Appendix Table D4).

3.3. Revealing spatial and temporal patterns of divergence

Next, we examine the spatial and temporal origins of the biases identified by our metric by conditioning likelihood distributions on patch center coordinates or local solar time (Section 2.3). This analysis reveals distinct spatial and temporal patterns in model errors. Figure 4 shows the divergence at each patch

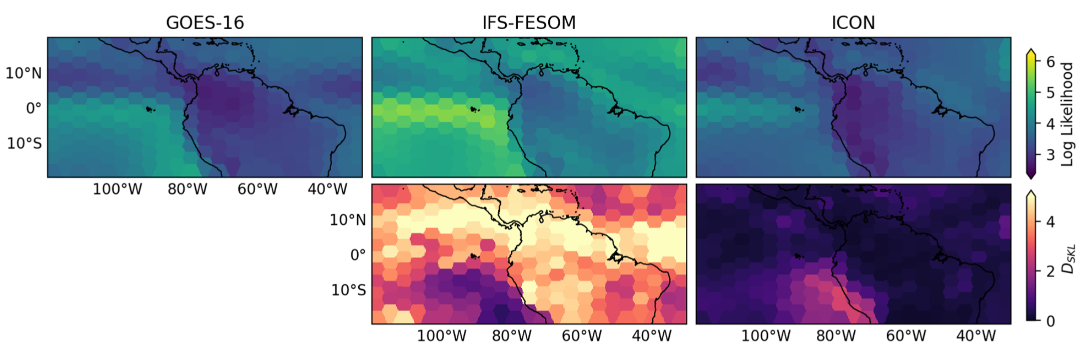

Figure 4. *Analysis of spatial biases in outgoing longwave radiation of two km-scale models, IFS-FESOM and ICON, compared to GOES-16 geostationary satellite observations. Top row: maps of mean log likelihood for each patch across the input region. Bottom row: maps of symmetrized KL divergence between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

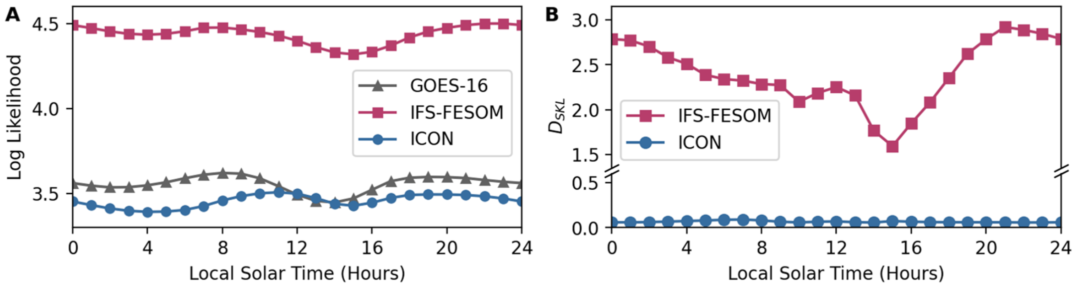

Figure 5. *Analysis of temporal biases of two km-scale climate models, ICON and IFS-FESOM, compared to GOES-16 geostationary satellite observations. Diurnal cycle of (A) average log likelihood and (B) the divergence between likelihood distributions of models and observations.*

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-FESOM, by comparison, exhibits high divergence more uniformly across the domain, with larger errors in convectively active regions, pointing to systematic biases in both cloudy and clear-sky regimes.

Temporal stratification reveals further structure in model biases (Figure 5). Deep convective clouds respond strongly to the diurnal cycle of incoming solar radiation, especially over land (Jones et al., 2023). Since clouds modulate OLR, their diurnal cycle is also expressed in the OLR signal. Climate models are known to struggle to capture the diurnal cycle accurately (Yin and Porporato, 2017), making it a critical test of model realism. The divergence between IFS-FESOM and GOES observations shows clear time-of-day dependence, with agreement improving in the early afternoon when convective activity peaks. While snapshot-based evaluation allows temporal analysis of biases, likelihood histograms converge after about 1 month (Supplementary Appendix Figure D5).

3.4. Metric comparison and sensitivity tests

We compare our method to relevant baselines: OLR MAE, SCWD, and multifractal parameters (Supplementary Appendix D). MAE reveals opposite biases to our approach, reflecting large-scale mean errors that are removed by histogram matching. SCWD closely mirrors MAE, indicating sensitivity to similar large-scale discrepancies. In contrast, multifractal analysis finds biases more similar to those revealed by our likelihood-based approach, consistent with both approaches probing fine-scale spatial variability. Our method is, however, more expressive, capturing model errors beyond scaling behavior alone. We further assess the sensitivity of our results to variations in NSF architecture, data standardization, and input patch size. Likelihood scores are consistent with those obtained for the original model setup. This robustness emphasizes the ability of our method to reliably evaluate models based on small-scale structural features.

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

We present a case study evaluation of two km-scale climate models, IFS-FESOM and ICON. Our results demonstrate that the likelihood-based method can robustly distinguish between models and observations, identifying spatio-temporally local biases in both models. Overall, ICON exhibits closer agreement with observations across regions and the diurnal cycle than IFS-FESOM. Nonetheless, it has a slight shift toward lower likelihoods compared with GOES, most likely because it simulates too small, unorganized convective cells, as expected for km-scale models with no convective parameterization (Wille et al., 2025; Takasuka et al., 2026). IFS-FESOM has an opposite consistent bias toward higher likelihoods, due to more organized convection and thus larger structures and clearer sky regions in OLR fields.

Our likelihood-based approach provides an objective, quantitative, and dataset-agnostic metric that captures both overall similarity and the spatial–temporal structure of model biases. The comparison with MAE and SCWD highlights that our framework provides a complementary perspective. While MAE and SCWD capture magnitude biases locally and in regional distributions, our method directly probes fine-scale spatial structures in simulated fields, offering sensitivity to physical processes such as cloud formation at resolutions relevant for km-scale models. For a comprehensive evaluation of model biases, traditional metrics need to be used alongside fine-scale focused methods such as the likelihood-based evaluation introduced here. Histogram matching removes differences in the OLR distribution, thereby discarding information about absolute magnitude biases. This choice is intentional to isolate morphological differences between models and observations. Alternative normalization strategies could be adopted if the evaluation objective is to retain sensitivity to absolute biases.

The framework introduced here enables rigorous comparison of simulations with observations, offering guidance for the calibration and improvement of next-generation kilometer-scale climate models. While our case study focused on outgoing longwave radiation, the framework is readily extensible. Future work includes incorporating additional variables, such as shortwave radiation, water vapor, or precipitation for a more comprehensive assessment of model realism.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/eds.2026.10054.

Acknowledgements. This work used supercomputing resources of the German Climate Computing Center (Deutsches Klimarechenzentrum, DKRZ) granted by its Scientific Steering Committee (WLA) under project ID 1153. We would also like to thank the anonymous reviewers for their constructive comments, who helped improve this manuscript.

Author contribution. Conceptualization: **L.J.F., T.R., R.C., P.S., H.M.C.** Methodology: **L.J.F., T.R., R.C., P.S., H.M.C.** Data curation: **L.J.F.** Data visualization: **L.J.F.** Writing original draft: **L.J.F.** All authors approved the final submitted draft.

Competing interests. The authors declare no competing interests.

Data availability statement. NextGEMS production simulations for ICON and IFS-FESOM 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 were 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. Model training and likelihood estimation code can be found on GitHub: https://github.com/lillif/likelihoods-km-scale-models. The repository has also been archived on Zenodo at DOI: 10.5281/zenodo.19358989.

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

Funding statement. L.J.F. acknowledges funding from the NERC Doctoral Training Partnership in Environmental Research Grant NE/S007474/1. H.M.C. was funded through the EERIE project (Grant Agreement No 101081383) funded by the European Union. University of Oxford's contribution to EERIE is funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (grant number 10049639). H.M.C. was also funded through a Leverhulme Trust Research Leadership Award "Seamless Uncertainty Quantification for Earth System prediction" (SUQCES). P.S. and T.R. acknowledge funding from the EU's Horizon Europe project Embed2Scale under grant agreement number 10113184. University of Oxford's contribution to Embed2Scale is funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (grant number 10113603). P.S. also acknowledges funding from the European Union's Horizon Europe project CleanCloud with grant agreement 101137639 University of Oxford's contribution to CleanCloud is funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (grant number 10113611). T.R. also acknowledges funding from ARIA and DSIT and Pillar VC under the Encode: AI for Science Fellowship. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Climate Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them. For the purpose of Open Access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

Provenance statement. This article is part of the Climate Informatics 2026 proceedings and was accepted in Environmental Data Science on the basis of the Climate Informatics peer review process.

References

Bony S, **Semie A**, **Kramer RJ**, **Soden B**, **Tompkins AM and Emanuel KA** (2020) Observed modulation of the tropical radiation budget by deep convective organization and lower-tropospheric stability. *AGU Advances 1*(3), e2019AV000155. https://doi.org/10.1029/2019AV000155.

Brunner L and Sippel S (2023) Identifying climate models based on their daily output using machine learning. *Environmental Data Science 2*, e22. https://doi.org/10.1017/eds.2023.23.

CIMSS (2025) *International Satellite Cloud Climatology Project — Next Generation (ISCCP-NG)*. Available at: https://cimss.ssec.wisc.edu/isccp-ng/ (accessed: 9th October 2025).

Dinh L, **Krueger D and Bengio Y** (2015) *NICE: Non-linear Independent Components Estimation* [arXiv:1410.8516 [cs.LG]]. https://doi.org/10.48550/arXiv.1410.8516.

Durkan C, **Bekasov A**, **Murray I and Papamakarios G** (2019) *Neural Spline Flows* [arXiv:1906.04032 [stat.ML]]. Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). https://doi.org/10.48550/arXiv.1906.04032.

Feng Z, **Leung LR**, **Liu N**, **Wang J**, **Houze RA**, **Li J**, **Hardin JC**, **Chen D and Guo J** (2021) A global high-resolution mesoscale convective system database using satellite-derived cloud tops, surface precipitation, and tracking. *Journal of Geophysical Research: Atmospheres 126*(8), e2020JD034202. https://doi.org/10.1029/2020JD034202.

Flato G, **Marotzke J**, **Abiodun B**, **Braconnot P**, **Chou SC**, **Collins W**, **Cox P**, **Driouech F**, **Emori S**, **Eyring V**, **Forest C**, **Gleckler P**, **Guilyardi E**, **Jakob C**, **Kattsov V**, **Reason C and Rummukainen M** (2013) Evaluation of Climate Models. In: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press, pp. 741–866.

Fraedrich K and Blender R (2003) Scaling of atmosphere and ocean temperature correlations in observations and climate models. *Physical Review Letters 90*(10), 108501. https://doi.org/10.1103/PhysRevLett.90.108501.

Fredriksen H-B and Rypdal K (2016) Spectral characteristics of instrumental and climate model surface temperatures. *Journal of Climate 29*(4), 1253–1268. https://doi.org/10.1175/JCLI-D-15-0457.1.

Freischem LJ, **Weiss P**, **Christensen HM and Stier P** (2024) Multifractal analysis for evaluating the representation of clouds in global kilometer-scale models. *Geophysical Research Letters 51*(20), e2024GL110124. https://doi.org/10.1029/2024GL110124.

Garrett RC, **Harris T**, **Li B and Wang Z** (2024) Validating Climate Models With Spherical Convolutional Wasserstein Distance [arXiv:2401.14657 [stat.AP]]. https://doi.org/10.48550/arXiv.2401.14657.

Gleckler PJ, **Taylor KE and Doutriaux C** (2008) Performance metrics for climate models. *Journal of Geophysical Research 113*, D06104.

Górski KM, **Hivon E**, **Banday AJ**, **Wandelt BD**, **Hansen FK**, **Reinecke M and Bartelmann M** (2005) Healpix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere. *The Astrophysical Journal 622*(2), 759–771. https://doi.org/10.1086/427976.

Hersbach H, **Bell B**, **Berrisford P**, **Hirahara S**, **Horányi A**, **Muñoz-Sabater J**, **Nicolas J**, **Peubey C**, **Radu R**, **Schepers D**, **Simmons A**, **Soci C**, **Abdalla S**, **Abellan X**, **Balsamo G**, **Bechtold P**, **Biavati G**, **Bidlot J**, **Bonavita M**, **De Chiara G**, **Dahlgren P**, **Dee D**, **Diamantakis M**, **Dragani R**, **Flemming J**, **Forbes R**, **Fuentes M**, **Geer A**, **Haimberger L**, **Healy S**, **Hogan RJ**, **Hólm E**, **Janisková M**, **Keeley S**, **Laloyaux P**, **Lopez P**, **Lupu C**, **Radnoti G**, **de Rosnay P**, **Rozum I**, **Vamborg F**, **Villaume S and Thépaut J-N** (2020) The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society 146*(730), 1999–2049. https://doi.org/10.1002/qj.3803.

Hohenegger C, **Korn P**, **Linardakis L**, **Redler R**, **Schnur R**, **Adamidis P**, **Bao J**, **Bastin S**, **Behravesh M**, **Bergemann M**, **Biercamp J**, **Bockelmann H**, **Brokopf R**, **Brüggemann N**, **Casaroli L**, **Chegini F**, **Datseris G**, **Esch M**, **George G**, **Giorgetta M**, **Gutjahr O**, **Haak H**, **Hanke M**, **Ilyina T**, **Jahns T**, **Jungclaus J**, **Kern M**, **Klocke D**, **Kluft L**, **Kölling T**, **Kornblueh L**, **Kosukhin S**, **Kroll C**, **Lee J**, **Mauritsen T**, **Mehlmann C**, **Mieslinger T**, **Naumann AK**, **Paccini L**, **Peinado A**, **Praturi DS**, **Putrasahan D**, **Rast S**, **Riddick T**, **Roeber N**, **Schmidt H**, **Schulzweida U**, **Schütte F**, **Segura H**, **Shevchenko R**, **Singh V**, **Specht M**, **Stephan CC**, **von Storch J-S**, **Vogel R**, **Wengel C**, **Winkler M**, **Ziemen F**, **Marotzke J and Stevens B** (2023) Icon-sapphire: Simulating the components of the earth system and their interactions at kilometer and subkilometer scales. *Geoscientific Model Development 16*(2), 779–811. https://doi.org/10.5194/gmd-16-779-2023.

Jones PW (1999) First- and second-order conservative remapping schemes for grids in spherical coordinates. *Monthly Weather Review 127*(9), 2204–2210. https://doi.org/10.1175/1520-0493(1999)127<2204:FASOCR>2.0.CO;2.

Jones WK, Stengel M and Stier P (2023) A lagrangian perspective on the lifecycle and cloud radiative effect of deep convective clouds over Africa. *EGUsphere 2023*, 1–25.

Judt F (2020) 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. https://doi.org/10.1175/JAS-D-19-0116.1.

Kingma DP and Dhariwal P (2018) Glow: Generative Flow With Invertible 1x1 Convolutions [arXiv:1807.03039 [stat.ML]]. https://doi.org/10.48550/arXiv.1807.03039.

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

Lee H-T, Laszlo I and Gruber A (2010) *ABI Earth Radiation Budget Upward Longwave Radiation: TOA* (Outgoing Longwave Radiation). *NOAA NESDIS Center for Satellite Applications and Research.* Available at: https://www.goes-r.gov/products/ATBDs/option2/RadBud_OLR_v2.0_no_color.pdf (Accessed 1st June 2026).

Lund R and Li B (2009) Revisiting climate region definitions via clustering. *Journal of Climate 22*(7), 1787–1800. https://doi.org/10.1175/2008JCLI2455.1.

Mooers G, Pritchard M, Beucler T, Srivastava P, Mangipudi H, Peng L, Gentine P and Mandt S (2023) Comparing storm resolving models and climates via unsupervised machine learning. *Scientific Reports 13*(1), 22365. https://doi.org/10.1038/s41598-023-49455-w.

Prein AF, Feng Z, Fiolleau T, Moon ZL, Núñez Ocasio KM, Kukulies J, Roca R, Varble AC, Rehbein A, Liu C, Ikeda K, Mu Y and Rasmussen RM (2024) Km-scale simulations of mesoscale convective systems over south america—a feature tracker intercomparison. *Journal of Geophysical Research: Atmospheres 129*(8), e2023JD040254. https://doi.org/10.1029/2023JD040254.

Rackow T, Pedruzo-Bagazgoitia X, Becker T, Milinski S, Sandu I, Aguridan R, Bechtold P, Beyer S, Bidlot J, Boussetta S, Deconinck W, Diamantakis M, Dueben P, Dutra E, Forbes R, Ghosh R, Goessling HF, Hadade I, Hegewald J, Jung T, Keeley S, Kluft L, Koldunov N, Koldunov A, Kölling T, Kousal J, Kühnlein C, Maciel P, Mogensen K, Quintino T, Polichtchouk I, Reuter B, Sármány D, Scholz P, Sidorenko D, Streffing J, Sützl B, Takasuka D, Tietsche S, Valentini M, Vannière B, Wedi N, Zampieri L and Ziemen F (2025) 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. https://doi.org/10.5194/gmd-18-33-2025.

Schmit TJ and Gunshor MM (2020) ABI Imagery from the GOES-R Series. In Goodman SJ, Schmit TJ, Daniels J and Redmon RJ (eds), *The GOES-R Series*. Amsterdam, The Netherlands: Elsevier, pp. 23–34. https://doi.org/10.1016/B978-0-12-814327-8.00004-4.

Schulzweida U (2023) CDO User Guide (Version 2.3.0). Zenodo. https://doi.org/10.5281/zenodo.10020800.

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

Stephens GL, Shiro KA, Hakuba MZ, Takahashi H, Pilewskie JA, Andrews T, Stubenrauch CJ and Wu L (2024) Tropical deep convection, cloud feedbacks and climate sensitivity. *Surveys in Geophysics 45*(6), 1903–1931. https://doi.org/10.1007/s10712-024-09831-1.

Stevens B, Satoh M, Auger L, Biercamp J, Bretherton CS, Chen X, Düben P, Judt F, Khairoutdinov M, Klocke D, Kodama C, Kornblueh L, Lin S-J, Neumann P, Putman WM, Röber N, Shibuya R, Vanniere B, Vidale PL, Wedi N and Zhou L (2019) Dyamond: the dynamics of the atmospheric general circulation modeled on non-hydrostatic domains. *Progress in Earth and Planetary Science 6*(1), 61. https://doi.org/10.1186/s40645-019-0304-z.

Takasuka D, Becker T and Bao J (2026) Precipitation characteristics and thermodynamic-convection coupling in global kilometer-scale simulations. *Journal of Advances in Modeling Earth Systems 18*(3), e2025MS005343. https://doi.org/10.1029/2025MS005343.

Tompkins AM and Semie AG (2017) Organization of tropical convection in low vertical wind shears: Role of updraft entrainment. *Journal of Advances in Modeling Earth Systems 9*(2), 1046–1068. https://doi.org/10.1002/2016MS000802.

Wille JD, Koch R, Becker T and Fischer E (2025) Extreme precipitation depiction in convection-permitting earth system models within the nextgems project. *Journal of Advances in Modeling Earth Systems 17*(7), e2024MS004840. https://doi.org/10.1029/2024MS004840.

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

Zhang X and Shao X (2015) Two sample inference for the second-order property of temporally dependent functional data. *Bernoulli 21*(2). https://doi.org/10.3150/13-BEJ592.

Cite this article: Freischem LJ, Reichelt T, Clark R, Stier P and Christensen H (2026). A generative likelihood framework for high-resolution climate model evaluation. *Environmental Data Science*, 5: e20. doi:10.1017/eds.2026.10054