Supplementary Material: A Generative Likelihood Framework for High-Resolution Climate Model Evaluation

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:

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

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. HEALPix Grid

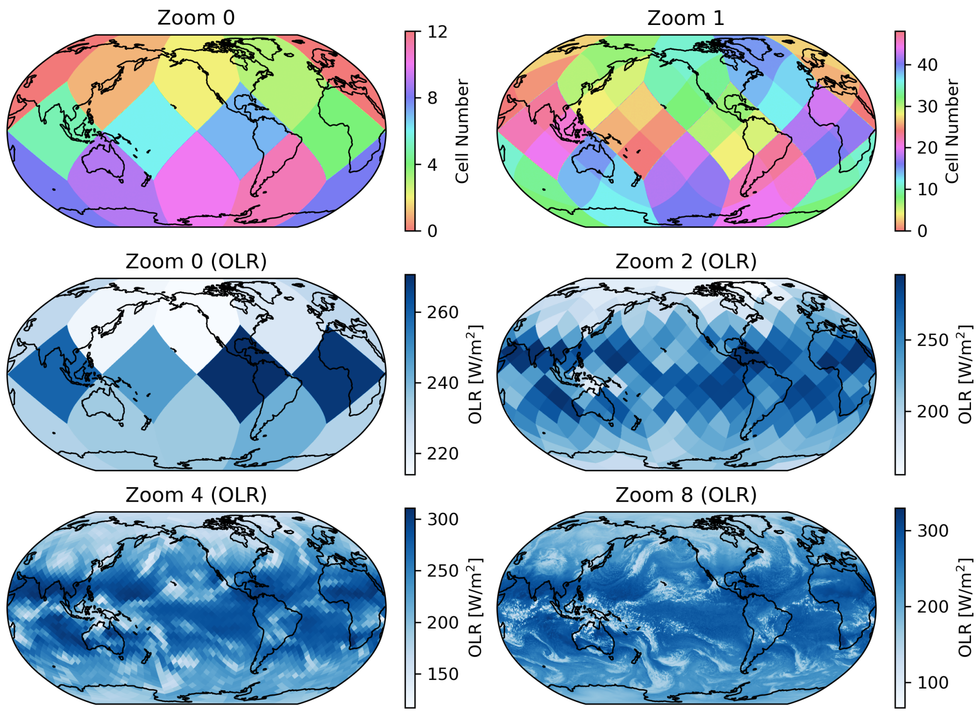

Figure B1. *Visualisation of the HEALPix projection. Top row: healpix cell numbers for zoom levels 0 and 1, Second and Third row: outgoing longwave radiation (OLR) on zoom levels 0, 2, 4, and 8.*

C. Neural Spline Flow Model of GOES-16 Observations

We model GOES-16 outgoing longwave radiation (OLR) patches using a Neural Spline Flow (NSF; Durkan et al., 2019). The base architecture is a multiscale normalising flow with 3 levels and 7 transformation steps per level. Each step begins with ActNorm, followed by affine coupling layers parameterised by rational quadratic splines with 4 bins and a tail bound of 1.0. Minimum bin width, height, and derivative constraints are set to 10^{-3}. The coupling networks are implemented as ResNets with 3 residual blocks, 96 hidden channels, batch normalisation, and no dropout. The model is trained for 20 epochs on a single NVIDIA A100 GPU using a batch size of 64.

To assess sensitivity to architectural choices, we trained two additional NSF variants: a smaller model with 3 levels and 5 steps per level using splines with 2 bins ("Small"), and a larger model with 4 levels and 7 steps per level using splines with 8 bins ("Large"). Validation-set likelihoods were consistent across configurations (Small: 3.51; Base: 3.53; Large: 3.52), indicating robustness to these hyperparameter choices.

The trained NSF generates realistic OLR patches (Figure C1) and closely reproduces the empirical OLR distribution of the GOES-16 training data (Figure C2).

To assess the model's suitability for efficient evaluation of large, high-resolution climate datasets, we measured its likelihood-estimation speed. Log-likelihoods for a batch of 64 (256) images are computed in 0.1 (0.25) seconds on average. Since likelihoods are calculated for individual patches independently, this process is easily parallelised.

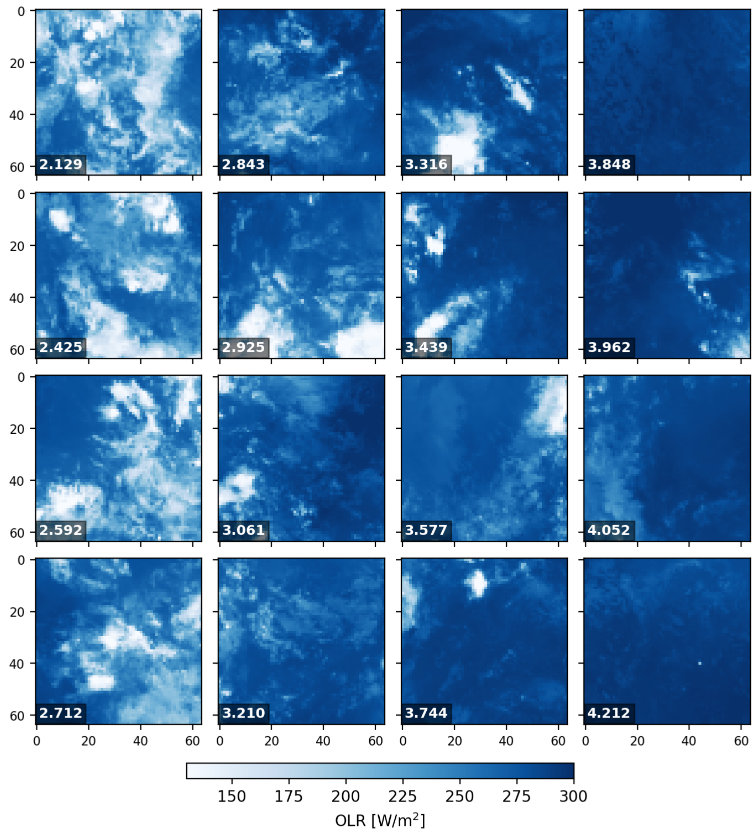

Figure C1. *Outgoing longwave radiation (OLR) images sampled from the Neural Spline Flow model trained on observations from the GOES-16 satellite. The first column shows low likelihood samples, columns two and three show random samples, and the fourth column shows high likelihood samples.*

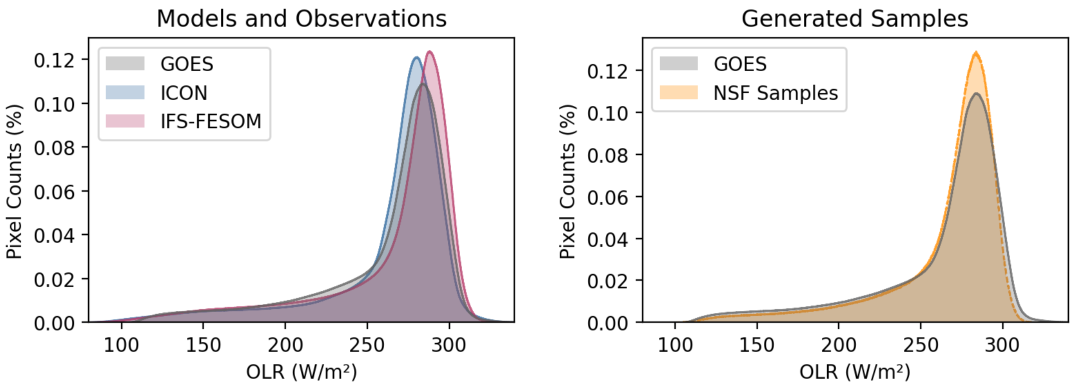

Figure C2. *Outgoing longwave radiation (OLR) histograms comparing (A) the GOES-16 satellite observations with the ICON and IFS-FESOM km-scale climate models and (B) the GOES-16 satellite observations with OLR samples generated by the trained Neural Spline Flow (NSF) model.*

D. Full similarity metric results

D.1. Baseline Climate Model Evaluation Metrics

D.1.1. Mean Absolute Error

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). For comparison with the patch-wise likelihood biases and multifractal biases (Table D2), we calculate the mean OLR for each patch location in the dataset, and compute MAE of each model as the difference between model OLR mean at that patch location compared to GOES OLR mean. 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. For a direct comparison with the spherical convolutional Wasserstein distance (Table D1), OLR MAE between models and GOES observations is calculated for individual pixels.

D.1.2. Spherical Convolutional Wasserstein Distance

We compare our metric against the spherical convolutional Wasserstein distance (SCWD) (Garrett et al., 2024) as a spatially aware distribution-based metric. To compute SCWD, we remap all datasets to a regular 0° lat–lon grid and use a linear convolutional kernel of size (6, 12). Due to the high computational cost, SCWD was evaluated only for the DJF months of 2024 (January, February, December). We tested the dependence of results to SCWD configurations by re-computing the distance for kernel sizes (3, 6) and (12,24) which changed the overall magnitude of the distances but not the ranking of distances across datasets or regional patterns.

Table D1. *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 the spherical convolutional Wasserstein distance (SCWD) all evaluated on OLR fields. Analysed were Janurary, February and December (DJF) of the year 2024 at four hourly resolution.*

	$D_{\mathrm{SKL}}(\mathcal{L}_{\mathrm{model}} \,\|\|\, \mathcal{L}_{\mathrm{obs}}) \downarrow$			**MAE** ↓			**SCWD** ↓		
	Overall	Ocean	Land	Overall	Ocean	Land	Overall	Ocean	Land
IFS-FESOM	2.351	2.522	5.054	11.124	11.583	9.787	4.207	4.390	3.673
ICON	0.114	0.200	0.036	14.695	14.522	15.200	5.601	5.882	4.784

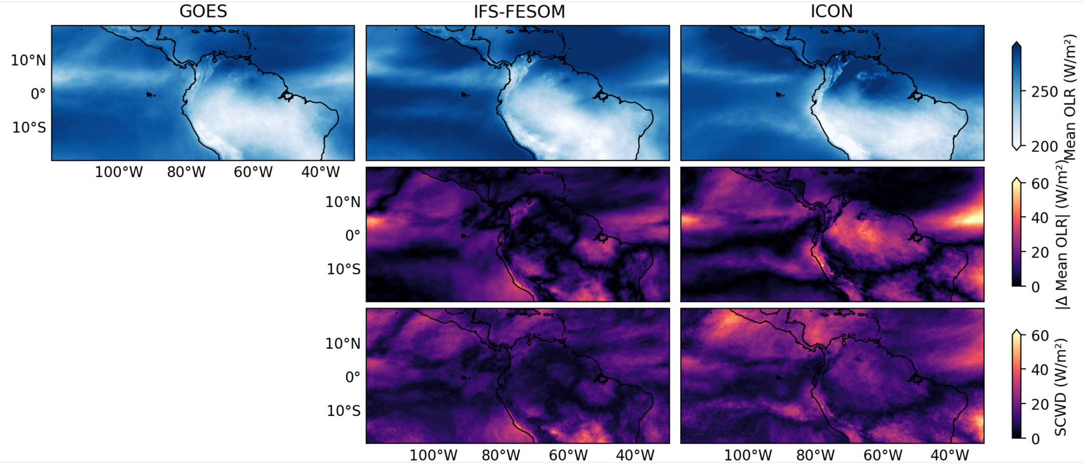

Figure D1. *Comparison of mean absolute error of outgoing longwave radiation (OLR) with the Spherical Convolutional Wasserstein Distance (SCWD). Analysed were January, February and December (DJF) of the year 2024 at four hourly resolution..*

D.1.3. Multifractal Parameter Bias

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 D2. *Biases identified by our likelihood based metric, $D_{\mathrm{SKL}}(\mathcal{L}_{\mathrm{model}}||\mathcal{L}_{\mathrm{obs}})$, compared to patch-wise mean absolute error (MAE$_{patch}$), 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$_{\text{patch}}$** $\downarrow$			**Multifractal** $\downarrow$		
	Overall	Ocean	Land	Overall	Ocean	Land	Overall	Ocean	Land
IFS-FESOM	1.496	2.432	2.332	4.680	4.436	5.258	1.633	1.309	2.402
ICON	0.057	0.110	0.008	7.950	7.465	9.100	1.040	1.190	0.684

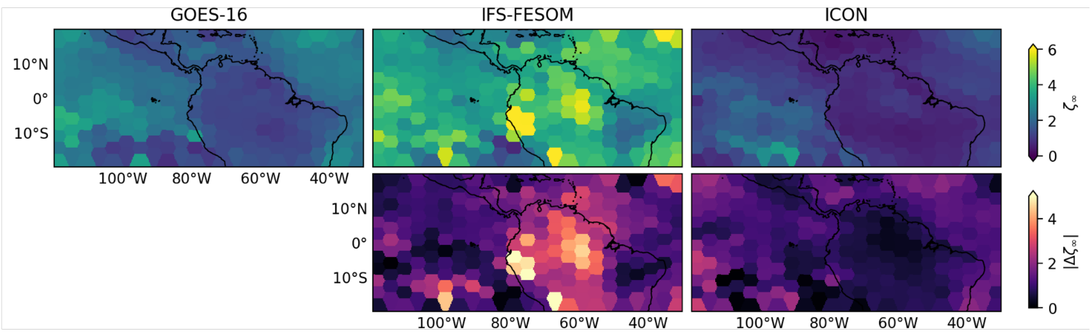

Figure D2. *Multifractal parameters (top) and bias compared to GOES (bottom) on an individual patch basis.*

D.2. Example Patches with Likelihoods

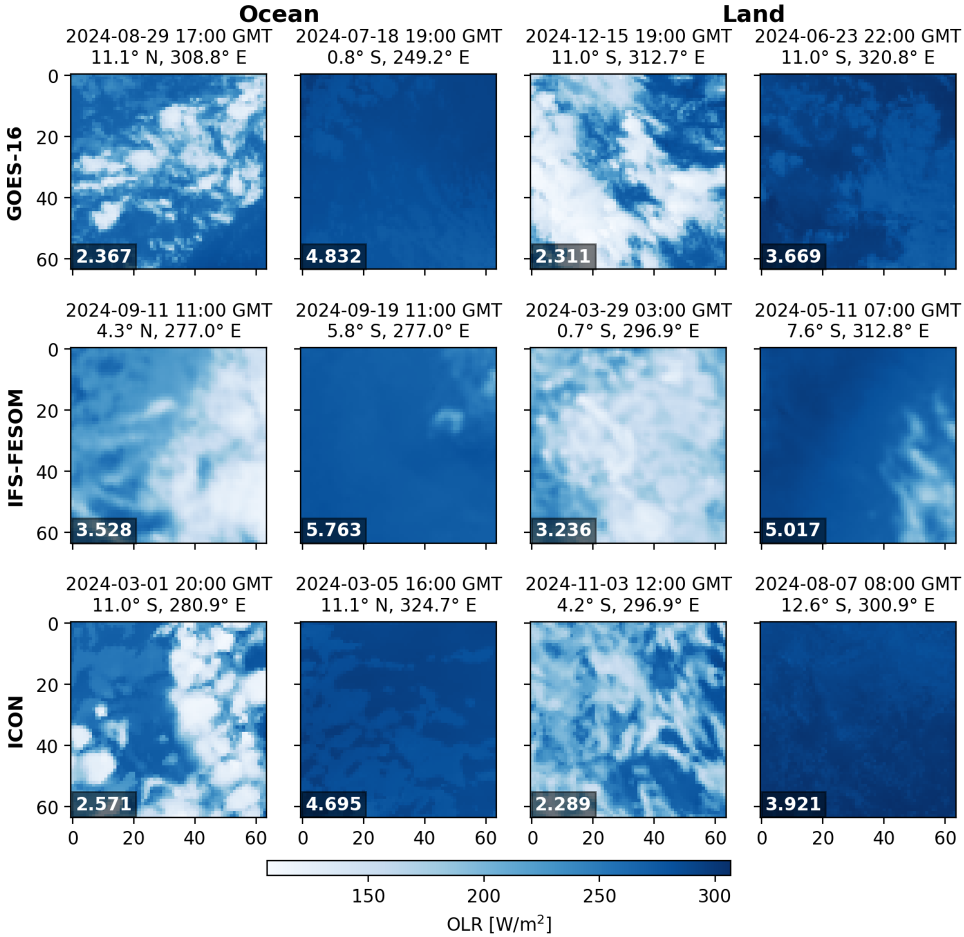

Figure D3. *Example patches with low and high log likelihoods from our three datasets: (top row) GOES satellite observations, (middle row) IFS-FESOM simulations and (bottom row) ICON simulations.*

D.3. Changing biases throughout the day

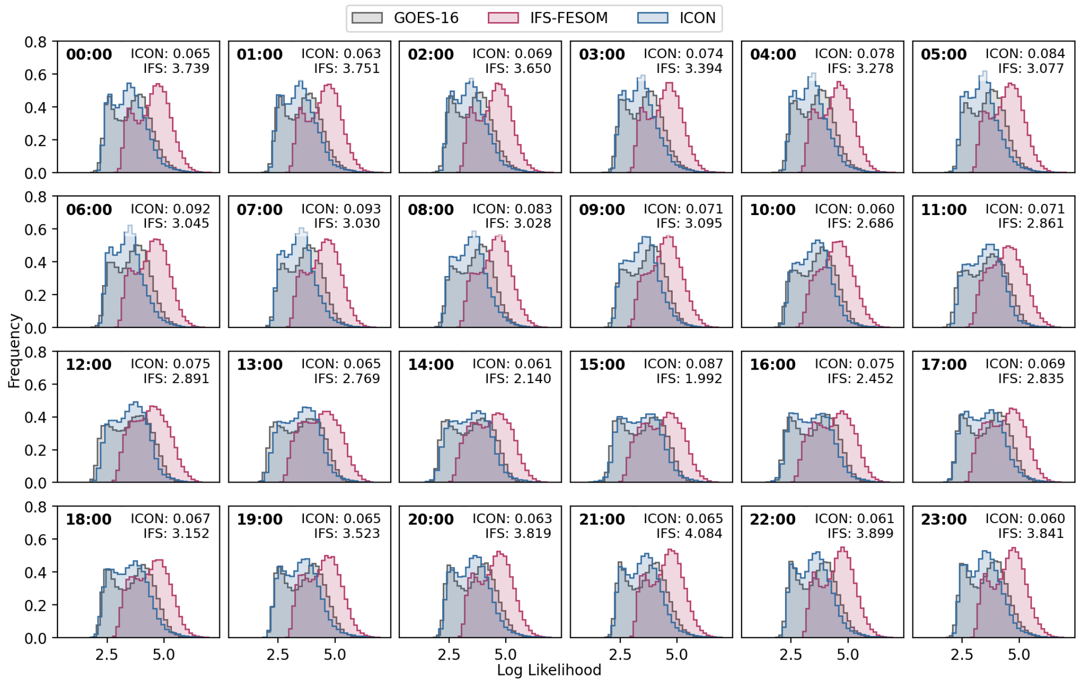

Figure D4. *Log likelihood distribution by local solar hour of GOES-16 geostationary satellite observations, compared to the two high-resolution climate models IFS-FESOM and ICON.*

D.4. Sensitivity Tests

Table D3. *Sensitivity of the symmetrised KL divergence to various hyperparameter choices: normalizing flow size, histogram matching (Hist. Match.), and patch size.*

Sensitivity Test		**Overall**	**Ocean**	**Land**
Small NSF	$D_{\mathrm{SKL}}(\mathcal{L}_{\mathrm{IFS-FESOM}} \,\|\|\, \mathcal{L}_{\mathrm{GOES}})$	1.400 ± 0.025	2.264 ± 0.030	2.145 ± 0.063
	$D_{\mathrm{SKL}}(\mathcal{L}_{\mathrm{ICON}} \,\|\|\, \mathcal{L}_{\mathrm{GOES}})$	0.056 ± 0.001	0.107 ± 0.001	0.009 ± 0.001
Large NSF	$D_{\mathrm{SKL}}(\mathcal{L}_{\mathrm{IFS-FESOM}} \,\|\|\, \mathcal{L}_{\mathrm{GOES}})$	1.476 ± 0.038	2.359 ± 0.061	2.294 ± 0.065
	$D_{\mathrm{SKL}}(\mathcal{L}_{\mathrm{ICON}} \,\|\|\, \mathcal{L}_{\mathrm{GOES}})$	0.057 ± 0.001	0.109 ± 0.001	0.014 ± 0.001
No Hist. Match.	$D_{\mathrm{SKL}}(\mathcal{L}_{\mathrm{IFS-FESOM}} \,\|\|\, \mathcal{L}_{\mathrm{GOES}})$	1.208 ± 0.020	2.257 ± 0.050	1.672 ± 0.046
	$D_{\mathrm{SKL}}(\mathcal{L}_{\mathrm{ICON}} \,\|\|\, \mathcal{L}_{\mathrm{GOES}})$	0.036 ± 0.001	0.067 ± 0.001	0.016 ± 0.001
32x32 Patches	$D_{\mathrm{SKL}}(\mathcal{L}_{\mathrm{IFS-FESOM}} \,\|\|\, \mathcal{L}_{\mathrm{GOES}})$	1.319 ± 0.074	1.935 ± 0.035	1.826 ± 0.084
	$D_{\mathrm{SKL}}(\mathcal{L}_{\mathrm{ICON}} \,\|\|\, \mathcal{L}_{\mathrm{GOES}})$	0.065 ± 0.001	0.124 ± 0.002	0.007 ± 0.001

D.5. Divergence within datasets

	Overall	Ocean	Land
$D_{\mathrm{SKL}}(\mathcal{L}_{\mathrm{GOES}_{1-15}} \,||\, \mathcal{L}_{\mathrm{GOES}_{16-31}})$	0.001	0.002	0.003
$D_{\mathrm{SKL}}(\mathcal{L}_{\mathrm{IFS-FESOM}_{1-15}} \,||\, \mathcal{L}_{\mathrm{IFS-FESOM}_{16-31}})$	0.003	0.004	0.002
$D_{\mathrm{SKL}}(\mathcal{L}_{\mathrm{ICON}_{1-15}} \,||\, \mathcal{L}_{\mathrm{ICON}_{16-31}})$	0.002	0.003	0.002

Table D4. *Symmetrised KL divergence of the likelihood distribution of outgoing longwave radiation fields of two km-scale climate models IFS-FESOM and ICON, compared to GOES-16 geostationary satellite observations.*

D.6. Variability in likelihood histograms

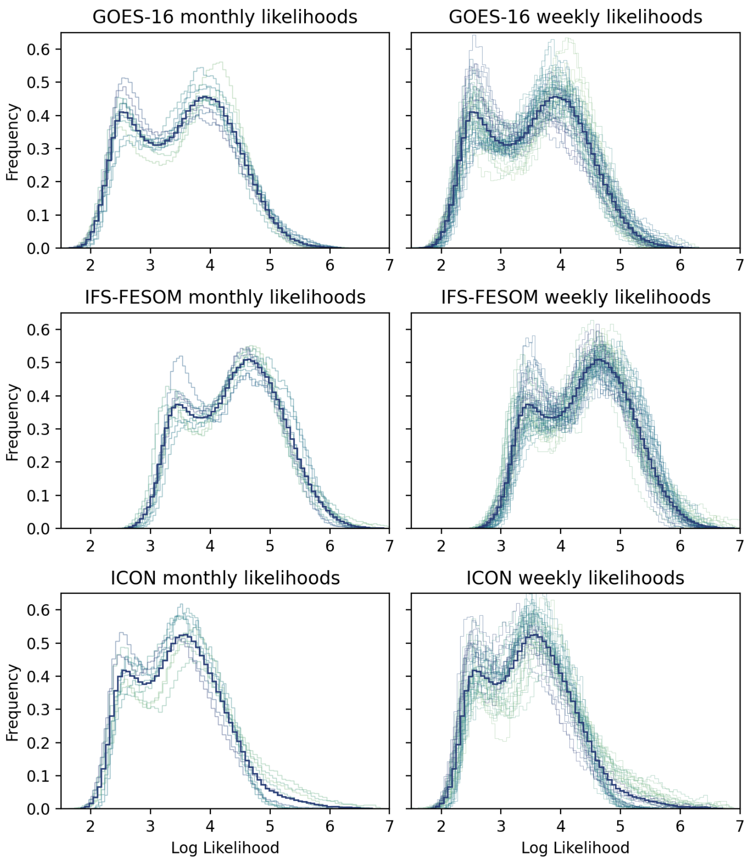

Figure D5. *Month-by-month (left) and week-by-week (right) variability of likelihood histograms of GOES-16 satellite observations and km-scale simulations of IFS-FESOM and ICON based on one year of data (2024). Thin lines are individual monthly (weekly) histograms and the thick line denotes the histogram for the entire year.*

References

Durkan, C., Bekasov, A., Murray, I., and Papamakarios, G. (2019). Neural spline flows.

Ellingson, R. G., Yanuk, D. J., Lee, H.-T., and Gruber, A. (1989). A technique for estimating outgoing longwave radiation from hirs radiance observations. *Journal of Atmospheric and Oceanic Technology*, 6(4):706 – 711.

Freischem, L. J., Weiss, P., Christensen, H. M., and Stier, P. (2024). Multifractal analysis for evaluating the representation of clouds in global kilometer-scale models. *Geophysical Research Letters*, 51(20):e2024GL110124.

Garrett, R. C., Harris, T., Li, B., and Wang, Z. (2024). Validating climate models with spherical convolutional wasserstein distance.

Gleckler, P. J., Taylor, K. E., and Doutriaux, C. (2008). Performance metrics for climate models. *Journal of Geophysical Research*, 113:D06104.

Lee, H.-T., Laszlo, I., and Gruber, A. (2010). ABI Earth Radiation Budget Upward Longwave Radiation: TOA (Outgoing Longwave Radiation).